THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Variational inference for Gaussian-jump processes with application in gene regulation

ANDREA OCONE

Doctor of Philosophy
Institute for Adaptive and Neural Computation
School of Informatics
University of Edinburgh
2013

Abstract

In the last decades, the explosion of data from quantitative techniques has revolutionised our understanding of biological processes. In this scenario, advanced statistical methods and algorithms are becoming fundamental to decipher the dynamics of biochemical mechanisms such those involved in the regulation of gene expression. Here we develop mechanistic models and approximate inference techniques to reverse engineer the dynamics of gene regulation, from mRNA and/or protein time series data.

We start from an existent variational framework for statistical inference in transcriptional networks. The framework is based on a continuous-time description of the mRNA dynamics in terms of stochastic differential equations, which are governed by latent switching variables representing the on/off activity of regulating transcription factors. The main contributions of this work are the following.

We speeded-up the variational inference algorithm by developing a method to compute a posterior approximate distribution over the latent variables using a constrained optimisation algorithm. In addition to computational benefits, this method enabled the extension to statistical inference in networks with a combinatorial model of regulation.

A limitation of this framework is the fact that inference is possible only in transcriptional networks with a single-layer architecture (where a single or couples of transcription factors regulate directly an arbitrary number of target genes). The second main contribution in this work is the extension of the inference framework to hierarchical structures, such as feed-forward loop.

In the last contribution we define a general structure for transcription-translation networks. This work is important since it provides a general statistical framework to model complex dynamics in gene regulatory networks. The framework is modular and scalable to realistically large systems with general architecture, thus representing a valuable alternative to traditional differential equation models.

All models are embedded in a Bayesian framework; inference is performed using a variational approach and compared to exact inference where possible. We apply the models to the study of different biological systems, from the metabolism in *E. coli* to the circadian clock in the picoalga *O. tauri*.

Acknowledgements

I would like to thank my supervisor Guido Sanguinetti, for all the time he dedicated to me. He guided and helped me through this work, which of course would not have been possible without him. I also thank him for his moral support, which converted all PhD worse moments in new strength and hopes. I was very lucky to be in his group and especially to have him as supervisor.

I want to thank Botond Cseke for having clarified a lot of conceptual and technical doubts about the work, and for having been (almost) always in the mood to answer to all my questions. Ten minutes with him were much more informative than a whole week with myself.

I am grateful to Manfred Opper and his group, especially Andreas Ruttor for having answered many emails about his exact inference code. I also thank Jeffrey Green and Matthew Rolfe, for insights about E. coli, and Andrew Millar for the collaboration which resulted in Chapter 5. I thank my advisors Jane Hillston and Subramanian Ramamoorthy for useful comments about the work and the schedule of my PhD. I thank my examiners, Ramon Grima and Mark Girolami, for having let me better contextualise this work by adding Chapter 6.

I would like to thank Gabriele Schweikert, Ronal Begg, Andrew Zammit Mangion and Botond, for having let me disturb their office during many afternoons. I thank people from the former group, Shahzad Asif, Grigorios Skolidis and Maurizio Filippone, for having shared with me good and bad moments of the PhD life. I thank people from the new group, Daniel Trejo, Anastasis Georgoulas and David Schnoerr, for nice coffee and lunch times. I thank all the people with which I shared my office, especially Davide Modolo and Dominik Grewe, and also people coming in the office, like Chris Fensch, for having lightened the days with random chats. A special thank to Alberto Magni, which spent with me most of the last weekends in the Forum, lunches followed by brownies and dinners with pizza. I also thank my flatmate, Rui Costa for our quick and funny conversations.

I am very thankful to Flavia (even if she doesn't want to be thanked), for her encouragement and for her love. Finally, I thank my family for supporting me and keeping all the time my morale up.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Andrea Ocone

Table of Contents

Chapter 1

Introduction

As molecular biology is becoming a more quantitative science, the mechanisms governing biological systems have begun to be a subject of extensive computational research. By supporting experimental results and producing *in silico* predictions, mathematical modelling aims to provide a deep understanding of biological systems. Thus, an increasing number of quantitative models are being developed and are becoming a fundamental component for research in molecular biology.

In this scenario, gene regulatory networks have received considerable attention from mathematical modellers. They represent sets of interacting genes, whose dynamical behaviour carries out crucial functions for the cell such as reproduction, metabolism, response to stimuli and so on. Thus, an understanding of the behaviour of gene regulatory networks would enable to control fundamental cellular mechanisms with a vast range of potential biomedical and biotechnological applications (Barnes et al., 2011).

The basic functional mechanism in gene regulatory networks is gene expression. Gene expression represents the process by which the genetic information encoded in the DNA is expressed into mRNA and then proteins. It involves several mechanisms, but the main steps are the following: transcription, where the gene is transcribed into mRNA, and translation, where mRNA is translated into protein (Fig. 1.1).

The expression of a gene is regulated by specific proteins called transcription factors, which are encoded by genes as well. Then, gene regulatory networks can be graphically represented as a set of nodes and edges: nodes represent the genes; edges represent the regulation of genes by protein products (i.e. transcription factors) of other genes.

Two main problems are related to gene regulatory networks: the first one is the network inference problem, which is the reconstruction of gene network architectures by using gene expression data or other data types (e.g. from microarray, sequencing technologies, etc). A large number of methods have been developed to solve the inference problem, ranging from regression analysis (Lèbre et al., 2010; Haury et al., 2012) to Bayesian networks (Friedman et al., 2000). An up-to-date evaluation of different methods can be found in (Marbach et al., 2012). Once a network structure is defined, the second related problem is to reconstruct the dynamics of the network. This essentially means to understand the mechanisms which regulate the evolution of mRNAs and proteins in gene regulatory networks. The present work concerns

this second issue.

A common approach to model gene expression dynamics is by using ordinary differential equations (ODEs). By denoting with $x(t)$ the quantity of interest at time t (e.g. mRNA or protein concentration), an ODE-based model is described by $\dot{x}(t) = f(x(t), \theta(t))$: the rate of change of $x(t)$ is a function $f(x(t), \theta)$, where θ represents some parameters (that can also be functions of time). As the mechanisms involved in gene expression have strong nonlinearities, a requirement for $f(x(t), \theta)$ is to be flexible enough to accommodate the behaviour of mRNA and protein time-courses. Therefore ODE-based models are usually highly parameterised and the function $f(x(t), \theta)$ is likely to be a nonlinear function.

In order to apply these models, we need a knowledge of the values of the parameters θ, such as the rate of production and degradation of mRNAs and proteins. Due to technical difficulties and cost, the experimental value of these parameters is usually not available; therefore, they have to be estimated from available data, such as mRNA or protein time-courses $x(t)$ (with $t = [0, T]$). Furthermore, usually the real concentrations $x(t)$ are unknown and the available data is just a corrupted version of $x(t)$. Then mathematical models are used to solve two related problems: 1) a state inference problem, which means to find the true concentrations $x(t)$ from their corrupted version; 2) a parameter learning problem, which means to estimate the value of the parameters θ of the model. A range of methods and tools for modelling gene regulatory networks are now available which allow us to simulate and study the properties of gene networks and give insights into different cellular processes (Vyshemirsky and Girolami, 2008b; Calderhead et al., 2009; Brunel and d'Alché Buc, 2010; Dondelinger et al., 2013).

In the present work we contribute to the research in this direction. We develop mechanistic models to reconstruct the dynamics of gene networks and solve the inference and learning problems using time-series gene expression data. The original methods developed here can be used to simulate/predict the dynamics of gene networks and may contribute to understand the molecular mechanisms underpinning cellular processes in different organisms. The methodology is presented in a general way, therefore it may also give a contribution to different areas from biology.

Two main principles drive the design of our methods. The first is the choice of the right level of abstraction for our models. This level should represent a trade-off between the simplicity needed for inference purposes and the flexibility needed to model complex gene expression dynamics. The second principle is the fact that an a priori knowledge of the biological system has to be considered. As we will describe in the next section, we will focus on the fact that gene expression is fundamentally a stochastic process and on the time-scale of different events involved in gene expression.

As it will be clear throughout the introduction and the rest of the thesis, these two principles make our models and methodologies original and substantially different from common ODE-based approaches.

In the following sections we first present the biological motivations for the models we use. Then we briefly describe our modelling and statistical approaches and report the work from literature which is related to our work. We conclude with an outline of the remaining chapters

and our contributions.

1.1 Biological background: the gene expression process

Our mathematical models of gene regulation are essentially driven by two important biological features: the mechanism of allostery and the stochasticity of gene expression. In the following subsections we briefly give a description of them.

1.1.1 Transcription and allosteric regulation

As we mentioned before, the transcription and translation mechanisms are the main steps to express a gene. Transcription involves the enzyme RNA polymerase (RNApol), which is responsible for the production of mRNA from the DNA sequence of the gene. Translation involves organelles called ribosomes, which synthesise proteins from mRNA. In reality, the whole process of gene expression is much more complex and can involve other mechanisms such as splicing (the removal of non-coding parts from mRNA) and transport of mRNA outside of the nucleus (present only in eukaryotes).

Here we try to give a general understanding of the gene expression process by focusing on transcription. We will not treat the translation mechanism.

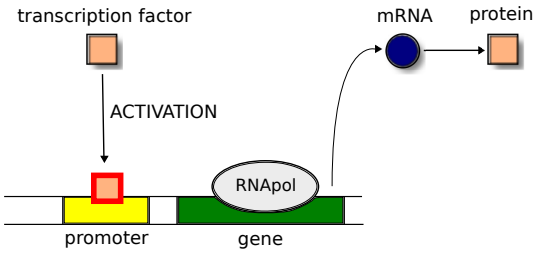

Figure 1.1: Representation of the gene expression process.

The enzyme RNApol is not specific for a single or a group of genes. Therefore, the expression of a particular gene depends essentially on the recruitment of RNApol to that gene. The recruitment of RNApol, known as regulation of gene expression, in turn depends on the presence of other proteins called transcription factors (TFs). TFs work by binding to the promoter of a gene, which is a region of the DNA upstream of the gene region, and recruiting RNApol for that gene. This is possible since TFs have two specific sites: one that is specific for the promoter, the other that is specific for RNApol. So different TFs work as adaptors between RNApol and different genes.

The regulation of gene expression is also a highly scheduled process, which is responsible to express specific genes with a specific timing. This timing essentially depends on the fact that TFs can be present in two different states: active and inactive. When they are in active state, they can bind to the promoter and trigger the transcription mechanism. On the other hand, when they are inactive, they cannot bind to the promoter and so the recruitment of RNApol does not occur.

	E. coli	*S. cerevisae*	Human fibroblast
Gene transcription	~ 1 min	~ 1 min	~ 30 min
Protein translation	~ 2 min	~ 2 min	~ 30 min
TF state transition	$1 - 100$ μs	$1 - 100$ μs	$1 - 100$ μs

Table 1.1: Time need to transcribe a gene, translate a protein and transit between TF states in bacteria (*E. coli*), yeast (*S. cerevisae*) and eukaryotic cells (human fibroblast) (Alon, 2006).

The activation/inactivation of TFs is often due to post-translational modifications regulated by allostery. Allostery (from Greek, other shape) is the mechanism by which TFs can change their three-dimensional shape and consequently have their function of adaptors (or not) between promoters and RNApol. In other words, the mechanism of allostery can switch on and switch off transcription by making the TFs active or inactive (Ptashne and Gann, 2002). Allosteric changes of TFs are in turn regulated by intra/extracellular signals.

To be precise, in absence of TFs, transcription occurs anyway but at low (basal) rate: RNApol can still spontaneously bind to the gene but with very low probability. In presence of an active TF, the transcription rate increases by a factor that depends on the efficiency of the TF to recruit RNApol.

In reality, there are two types of TFs: transcriptional activators or transcriptional repressor. When they bind to the promoter of a gene, activators recruit RNApol, while repressors prevent RNApol to start the transcription. Therefore there are two possible mechanisms to increase the transcription rate of a gene: the binding of an activator to the promoter (activation) or the unbinding of a repressor from the promoter (de-repression).

To incorporate this knowledge into a dynamical system, it is necessary to give an idea of the time scales of the different mechanisms involved in the regulation of gene expression. In particular we emphasise the difference in the timing between transcriptional (and translational) mechanisms and the post-translational mechanism (by which TFs change their activity state). As shown in Table 1.1, the ratios between the time needed for transcription/translation and the time needed for the TF to change state are in the order: $10^2 - 10^4$ in bacteria and yeast, $10^4 - 10^6$ in eukaryotes (Alon, 2006).

1.1.2 Stochasticity in gene expression

The mechanisms of transcription and translation involve random biochemical reactions, such as the binding of the TF to the promoter and the binding of the RNApol to the gene. As a consequence, gene expression is essentially a stochastic process (Ozbudak et al., 2002): the number of mRNAs and proteins produced during time from a gene, can be described by a deterministic component and unpredictable fluctuations (i.e. noise) around this component[1].

Two sources of noise can be distinguished in the stochastic fluctuations of gene expression: an intrinsic noise component and an extrinsic noise component (Swain et al., 2002; Elowitz et al., 2002). The former originates from the intrinsically random biochemical reactions, while

[1]By writing this, we do not mean that the noise is purely additive (Shahrezaei et al., 2008).

extrinsic noise depends on the "state" of the cell. This state is represented by the components of the gene expression machinery, such as the number of TFs, RNApol and ribosomes, and other features such as cell cycle stage, that are variable from cell to cell.

Since the noise propagates in gene networks (Pedraza and van Oudenaarden, 2005), the stochastic component plays a fundamental role in the dynamic interactions between genes (Eldar and Elowitz, 2010). Thus, some studies try to understand how the noise component affects the gene expression from the simple regulatory interaction to the level of gene regulatory networks (Macneil and Walhout, 2011; Chalancon et al., 2012).

While there is increasing interest in understanding how important cellular functions, such as metabolism and stress response, critically depend on the presence of the noise (Süel et al., 2006; Acar et al., 2008), in this thesis we focus on the role of stochasticity from a statistical inference perspective.

1.2 Models and methods

1.2.1 Modelling allosteric regulation

One of the main features that has driven the design of our mathematical models is the mechanism of allostery. As we have described above, allostery can switch on and off the transcription of a gene by changing the activity state of the TFs. From a quantitative perspective this mechanism represents a discrete on/off model: transcription is turned on when the TF is in active state; transcription is turned off when the TF is in inactive state. On the other hand, the gene expression level (e.g. concentration of mRNA) will range in a continuous fashion[2] between a minimum value, which is determined by the basal rate at which transcription occurs, and a maximum value, which can be reached only if transcription is activated by the TF[3].

From a mathematical point of view, the whole mechanism of gene expression can be intuitively expressed as an hybrid on/off model, composed of discrete variables (i.e. TF states) and continuous variables (i.e. gene expression levels). In this model, TFs have a discrete binary state which can switch between two values, active or inactive[4], and gene expression levels change continuously within a lower and an upper boundary. This means that, while gene expression levels can change their values as a continuous function, TFs change their activity only through jumps from one discrete state to the other. Such a feature is crucial in our modelling assumption: it reflects the fact that transitions between TF states occur in a shorter time scale compared to transcription/translation mechanisms.

[2] The continuous assumption is valid when the number of individuals (e.g. mRNA molecules) is rather large.

[3] If the TF is a repressor, then we have an opposite behaviour: the maximum gene expression level depends on the basal transcriptional rate, while the minimum is reached when the TF is in active state.

[4] As we will present in Chapter 5, an alternative interpretation of the on/off mechanism is the following. Instead of modelling a binary TF state, we can model a binary promoter state: the promoter can be occupied or not by the TF. This interpretation is equivalent to the previous one, if we assume that every time the TF is in active state, then it occupies the promoter and starts transcription.

1.2.2 Modelling stochastic effects

In order to incorporate stochastic effects in gene regulation models, in this thesis we use a mathematical framework based on stochastic continuous-time models. Stochastic continuous-time models have a long tradition in many research areas, from statistical physics (Gardiner, 2009) to econometrics (Preis et al., 2011) and ecology (Renshaw, 1991). As they are able to handle naturally the noise intrinsic in a physical system, they represent an attractive framework to model the stochasticity of biochemical reactions involved in gene expression. Moreover, there are several advantages in the use of continuous-time models compared to discrete-time models. First, the presence of noise at continuous times enables a number of dynamical behaviour such as bistability (Gammaitoni et al., 1998), noisy limit cycles and quasi-cycles (Wallace et al., 2011), which are not possible if noise is injected in the system only at discrete times. In addition, by working in continuous-time we bypass the problem of setting the sampling rate which instead is needed if we work with discrete-time models.

Mathematically, stochastic continuous time-models can be described in terms of stochastic differential equations (SDEs). While a representation of the intrinsic component of gene expression is inherent in the definition of SDE (e.g. as a Wiener process), it is not obvious how to model the extrinsic noise component. In this work, we encode the extrinsic noise in the state of the TF, by modelling its activity through a stochastic process (e.g. a 2-state Markov jump process). Thus, extrinsic noise sources due to the actual state of the cell can be taken into account.

The presence of noise in gene expression means that such a process can be studied only through a statistical description[5]. Ergo, the quantities involved in the model such as TF states and gene expression levels must be expressed in terms of probability distributions.

The data we use in this work are time-series of mRNA/protein concentration produced from the expression of several genes. Since these measurements are obtained from cell populations, the intrinsic and extrinsic stochasticity of gene expression should be averaged out. Therefore, the choice of the mathematical models we use in this thesis could be object of criticism. One can argue that the models are selected on the basis of mathematical convenience[6] rather than as a reflection of any biological knowledge. In any case, the fact that we are using SDE-based models means that our models could theoretically be used on a wider class of time-series data (e.g. gene expression data measured at the single cell level).

1.2.3 Statistical methods

When we have to deal with real data, another source of noise (in addition to stochasticity) has to be considered: the noise given by the measurement error. In this thesis, we cope with the uncertainty over the data by embedding the inference and learning problems in a Bayesian framework. This means that the results are obtained in terms of posterior probability distri-

[5]Given its statistical nature, the subject of this thesis belongs to the area of statistical systems biology (Stumpf et al., 2011) rather than generic computational systems biology (Kitano, 2002).

[6]We will see in Chapter 5 that a mean field approximation is only possible when we consider noise in the system.

butions, which take into account of two fundamental aspects: the fact that data are noisy and that these data are generated by stochastic continuous-time models with some prior knowledge about the biological mechanisms. Once the posterior probability distributions have been obtained, they can be used to find statistics (e.g. mean values and levels of uncertainty) about the quantities of interest.

The main problem of Bayesian statistics is that the computational cost required for the posterior distribution can become prohibitively high when we consider systems with increasing size. However, it is still possible to compute approximations to the intractable posterior, using methods of essentially two classes: deterministic, such as Laplace and variational approximations, and stochastic, such as Markov chain Monte Carlo methods (MCMC).

In this work we will focus on the first class of methods, especially on variational approaches. Usually, it is not straightforward how to derive variational approximations to the posterior which are at the same time accurate and tractable. However, the main advantage of variational approximations over sampling methods is the reduced computational cost. Therefore they can potentially be used for large systems, such as gene regulatory networks, without suffering from the "curse of dimensionality".

1.3 Related work

There is a vast literature about using stochastic continuous-time models to learn the parameters in gene networks. These models have been extensively developed in the area of financial econometrics (Johannes and Polson, 2003) and are now expanding towards biological applications. Methods are mostly based on MCMC and include: maximum likelihood approaches (Reinker et al., 2006; Tian et al., 2007), fully Bayesian approaches (Boys et al., 2008; Golightly and Wilkinson, 2006), approximate Bayesian computation schemes (Toni et al., 2009) and moment matching methods (Zechner et al., 2012; Kügler, 2012).

Among all these methods, we briefly mention the work of Stimberg et al. (Stimberg et al., 2011, 2012), which is closely related to our research. Stimberg and colleagues use the same continuous-time representation of transcriptional regulation that we use (a conditionally Ornstein-Uhlenbeck process) and solve the inference and learning problems by developing sampling based methods. They develop a change point model for the switching dynamics of the TF states (which is modelled as a 2-state Markov jump process), where also the system noise can switch in a bistable mode. These switchings are completely defined by: a set of parameters, representing the values of the states, and a set of positions, representing the times at which the jumps occur. As the jumps are considered as Poisson events, the distribution of the inter-jump intervals Δt (i.e. the intervals between events) is exponentially distributed: $p(\Delta t) = f \exp(-f\,\Delta t)$, where f represents the jump rate (i.e. the expected number of jumps per unit time). By using a Gaussian likelihood and the fact that the Ornstein-Uhlenbeck process is a Gaussian process, they compute analytically the marginal posterior over the 2-state Markov jump process. The marginal posterior is then used to draw samples: they implement a Metropolis-within-Gibbs sampler, alternating between sampling the set of parameters and the position of the change

points.

A vast literature is also present for inference in gene network dynamics using ODE-based models. We mention the work of Lawrence et al. (Lawrence et al., 2007), where the latent TF dynamics is modelled by a function $g(f)$, where f is drawn from a Gaussian process. They solve the inference and learning problem in two different cases: a linear case, where $g(\cdot)$ is linear, and a nonlinear case. In the linear case, they show that the gene expression is given by a Gaussian process. By choosing an exponential radial basis function kernel, they compute analytically the mean and covariance of the posterior Gaussian process on $g(f)$. Parameters are estimated by using a type-II maximum likelihood. In the nonlinear case, they are still able to find a maximum a posteriori solution, using a Laplace approximation. For the nonlinear case, a Markov chain Monte Carlo algorithm to sample from the posterior process was also derived by Titsias et al. (Titsias et al., 2009).

There is a crucial difference between models based on a Gaussian process prior and models based on Markov jump process prior. In brief, the formers do not allow rapid transitions in the TF activity and so they cannot capture the characteristic switching behaviour given by allosteric regulation. On the other hand, models based on Markov jump processes are intrinsically switching models.

The scientific literature is also rich of other methods to infer the dynamics of gene regulation, which are based on discrete approximations (Barenco et al., 2006; Rogers et al., 2007) (e.g. piecewise linear approximations of the TF dynamics). These methods cannot provide the subtleties of a continuous-time reconstruction of the gene expression dynamics, so they are not much relevant to this thesis.

1.3.1 Approximations of the chemical master equation

Stochastic chemical kinetics provides a foundation for describing biochemical dynamics in terms of Markov jump process (Wilkinson, 2011). As Markov jump processes are defined through the chemical master equation (CME) (Gardiner, 2009), this equation can be successfully employed to model gene network dynamics.

The advantage of a CME description is that both deterministic and stochastic components of a system arise directly from this equation. In other words, the evolution of a system simulated by using the CME, consists of both a deterministic (macroscopic) and a stochastic (microscopic) component.

Exact statistical inference in systems described in terms of CME is possible, but it becomes an intractable problem when the number of (molecular) species is large (Ruttor et al., 2010). Opper and Sanguinetti (Opper and Sanguinetti, 2008) provided a tractable solution by defining the problem in a variational framework and using a mean field approximation. They replace the posterior joint probability over the discrete paths of all (molecular) species with a factorised distribution and obtain an iterative algorithm to perform inference and parameter estimation with low computational cost.

Alternative approaches can be obtained by directly approximating the CME as a partial dif-

ferential equation (PDE). This is possible by treating the deterministic and stochastic components (intrinsically coupled in the CME) as independent, under the assumption that the system size is large enough.

Two possible approximations of the CME have been made: the first is obtained by second-order Taylor expanding the CME (Kramers-Moyal expansion (Risken, 1984)) and discarding the second-order terms. The resulting equation is a nonlinear Fokker-Planck equation representing a diffusion process[7]. The other approximation was obtained by Van Kampen (Van Kampen, 1981), by expanding the CME as a function of powers of the system volume variable Ω. In the limit of large Ω, terms of order $\mathcal{O}\left(\Omega^{-1}\right)$ and higher can be neglected; the resulting approximation, commonly known as linear noise approximation (LNA), is a linear Fokker-Planck equation.

Both of these approximations have been used in a number of works to solve the problem of parameter estimation in gene networks. Golightly et al. (Golightly and Wilkinson, 2005, 2011) use the CLE in a fully Bayesian approach to estimate the parameters of a stochastic continuous-time model which is observed at discrete times. They adopt two different inference strategies based on sampling: a MCMC method based on data augmentation (previously used on stochastic volatility models in a mathematical finance context) (Golightly and Wilkinson, 2005) and a particle MCMC method (Golightly and Wilkinson, 2011).

Stathopoulos et al. (Stathopoulos and Girolami, 2012) solve the same parameter estimation problem, by proposing a LNA and applying Riemann manifold MCMC methods (Girolami and Calderhead, 2011). Also Komorowski et al. (Komorowski et al., 2010) use a LNA to write explicitly the likelihood of the data, which is then used in a Bayesian approach to draw samples with a Metropolis-Hastings algorithm.

A different approach is developed by Ruttor et al. (Ruttor and Opper, 2009). They assume that the evolution of molecule numbers can be described by Gaussian fluctuations around a deterministic state. By using a Van Kampen's system size expansion (which they call *weak noise approximation*), they derive backward and forward ODEs for the moments of the Gaussian variables, which are used to solve the inference problem without the need of MCMC methods.

1.4 Structure and contributions

The work in this thesis develops ideas regarding quantitative modelling of gene network dynamics using hybrid stochastic continuous-time models. Our central focus is on the problems of statistical inference and parameter estimation, that in such models is non-trivial.

The work is developed in five main chapters: the first one describe the concepts and methodologies used in this research. The remaining four chapters represent the three main contributions and a comparison with existing inference methods. Below we report the content of each chapter, by emphasising the main contributions.

[7] In a different way, Gillespie (Gillespie, 2000) derived a chemical Langevin equation (CLE) and showed that it is associated with the same nonlinear Fokker-Planck equation. How accurate is the CLE with respect to the CME is a current research subject (Grima et al., 2011).

- **Chapter 2** introduces the background material that is needed to understand the remaining chapters. It starts with a definition of stochastic process and stochastic differential equation, using concepts from probability theory. The introduction of stochastic processes is essential to describe physical dynamical systems such as the ones we are treating in this thesis. Then, we briefly introduce the Bayesian approach, which represents the probabilistic way we deal with prior knowledge and (noisy) data from lab experiments. We show that probabilistic models can be easily represented by using graphical models and that the introduction of latent variables provides a tractable way to model the complexity of dynamical systems. By using this latent variable description, we show how to solve the state inference and parameter estimation problems of a continuous-time dynamical system which is discretely observed. Finally we introduce approximate inference methods, which provide an efficient way to solve the inference and learning problems.

- **Chapter 3** introduces Gaussian-jump processes and present two approaches to inference in partially observed Gaussian-jump processes. One of the approaches is an exact inference method (Sanguinetti et al., 2009); the other is an approximate inference method which is developed in a variational framework (Sanguinetti et al., 2009). We focus on the approximate method and show how the inference problem can be turned into an optimisation problem. By using a conditional approximation and restricting our interest to the deterministic case (where system noise is zero), we show how to solve the optimisation problem through a constrained optimisation algorithm.

  This algorithm, based on functional gradients, represents the first main contribution of this thesis. It provides an excellent approximation, comparable to results from the exact inference method, and with a very contained computational cost. The method can be easily generalised to the case where multiple Gaussian-jump processes[8] are driven by multiple interacting jump processes (with pairwise interactions).

  In the rest of the chapter we first show the comparison on simulated data of the exact and variational methods. Then we report an application of the variational inference approach on a study of *E. coli*'s metabolism, which contributed to the work of Rolfe et al. (Rolfe et al., 2012).

- **Chapter 4** is a paper by Ocone and Sanguinetti (Ocone and Sanguinetti, 2011). Here we develop statistical inference models for hierarchical networks such as feed-forward loops. These models allow to perform inference in structures that are not limited to the single-input motif, but include an intermediate layer of regulation.

  The hierarchical model is based on the Gaussian-jump process representation described in Chapter 3 (in its deterministic version), where we use a Heaviside step function to model the switching activity of the regulator in the intermediate layer. From a methodological point of view, we extend the constrained optimisation algorithm mentioned

[8] In reality, when the system noise is zero, the Gaussian-jump process becomes a system which, conditioned on the states of the jumps, is deterministic.

above. This is done by introducing a Laplace-type approximation to compute additional moments due to the intermediate layer of regulation.

We show an application of the methodology on two real data sets: one including the *p53* gene, the other including a feed-forward loop involved in *E. coli*'s metabolism.

- **Chapter 5** is a paper by Ocone et al. (Ocone et al., 2013). We present a framework to model gene regulatory networks. The framework is based on a representation of the transcriptional regulation as a Gaussian-jump process, as in (Opper et al., 2010). We introduce state-dependent jump rates to enable the modelling of multiple Gaussian-jump processes connected in an arbitrary structure.

  Inference is obtained with two methods. We extend the exact inference method in (Sanguinetti et al., 2009) to the general case of state-dependent jump rates. Then we adopt a variational approach as in (Opper et al., 2010). Here a conditionally Ornstein-Uhlenbeck process description of the gene expression mechanism, enables the use of a mean field approximation. In this approximation, the variational distribution factorises into the product of pure Gaussian and pure jump processes. The variational approach turns the inference problem in an optimisation problem which is solved by combining two strategies. A fast forward-backward procedure is used to optimise the jump process component of the mean field approximation, while a constrained gradient descent algorithm based on functional derivatives is used to optimise the Gaussian process component.

  The framework allows to model complex dynamics of (potentially large scale) gene regulatory networks and provide comparable or better predictions than alternative methods. We apply the methodology to two real data sets: one from a synthetic gene network in the yeast *Saccharomyces cerevisiae* and another from a study of the circadian clock in the picoalga *Ostreococcus tauri*.

- **Chapter 6** describes alternative methods to perform inference from discrete data. This is useful to put the work presented in this thesis in the context of other existing methods.

Part of the work in this thesis has been published in the following articles

- Ocone A, Sanguinetti G (2011): Reconstructing transcription factor activities in hierarchical transcription network motifs. *Bioinformatics* 27(20): 2873-9

- Rolfe MD, Ocone A, Stapleton MR, Hall S, Trotter EW, Poole RK, Sanguinetti G, Green J (2012): Systems analysis of transcription factor activities in environments with stable and dynamic oxygen concentrations. *Open Biology* 2(7): 120091

- Ocone A, Sanguinetti G (2013): A hybrid statistical model of a biological filter. *Proceedings Third International Workshop on Hybrid Autonomous Systems* EPTCS 124: 100-108

- Ocone A, Millar AJ, Sanguinetti G (2013): Hybrid regulatory models: a statistically tractable approach to model regulatory network dynamics. *Bioinformatics* 29(7): 910-6.

Chapter 2

Methods

We briefly introduce background material which is useful to understand the rest of the thesis. The chapter is divided in five main sections. In Section 2.1 we start by defining stochastic processes to describe probabilistic dynamical systems. We focus on Markov processes and derive the Chapman-Kolmogorov equation. Then we introduce stochastic differential equations and from these we derive equations for the moments in a linear and a general case.

In Second 2.2 we introduce the Bayesian framework, as a way to get information about dynamical systems from noisy measurements. We describe the maximum likelihood and maximum a posteriori estimates.

In Section 2.3 we describe graphical models, which are useful to give graphical representations of probabilistic models. We focus on Markov models and then introduce the state space representation.

In Section 2.4 we look at a system described by a continuous-time model where we have access only to discrete-time noisy observations. From these observations we try to obtain information about the state of the system (state inference problem) and to estimate the parameters of the model (parameter learning problem).

Finally, in Section 2.5 we introduce approximate inference methods which are needed when inference and learning problems are intractable.

2.1 Stochastic processes

Probabilistic systems which evolve in time can be described in terms of stochastic processes. A stochastic process is a collection of random variables $\mathbf{X_1}, \mathbf{X_2}, \ldots$ indexed by time $\mathbf{t} = \{t_1, t_2, \ldots\}$. By considering a particular possible random variable $\mathbf{X}$ of the stochastic process, the ordinary function of time $\mathbf{X(t)}$ is called realisation, trajectory or sample path of the stochastic process (Van Kampen, 1981). According to the discrete or continuous nature of the time variable $\mathbf{t}$, stochastic processes can be classified as discrete-time or continuous-time stochastic processes. Here we will focus on continuous-time stochastic processes, therefore sample paths are considered as infinite dimensional objects.

2.1.1 Markov property

Assuming we can measure the value $\boldsymbol{x}$ of the random variable $\mathbf{X}$ at each time point t, the system can be completely described by the joint probability density[1] $p(\boldsymbol{x}_1, t_1; \boldsymbol{x}_2, t_2; \ldots; \boldsymbol{x}_n, t_n)$. Since the probabilistic system evolves in time, we are interested in predicting the future value of $\mathbf{X}(\mathbf{t})$. This is done by defining a conditional probability density

$$p(\boldsymbol{x}_{n+1}, t_{n+1} | \boldsymbol{x}_1, t_1; \boldsymbol{x}_2, t_2; \ldots; \boldsymbol{x}_n, t_n) = \frac{p(\boldsymbol{x}_1, t_1; \boldsymbol{x}_2, t_2; \ldots; \boldsymbol{x}_n, t_n; \boldsymbol{x}_{n+1}, t_{n+1})}{p(\boldsymbol{x}_1, t_1; \boldsymbol{x}_2, t_2; \ldots; \boldsymbol{x}_n, t_n)} ,$$

which determines the future value $X_{n+1} = \mathbf{X}(t_{n+1})$, given the knowledge of all the past values $X_1, X_2, \ldots, X_n$. When only the present value X_n is needed to determine the future value X_{n+1},

$$p(\boldsymbol{x}_{n+1}, t_{n+1} | \boldsymbol{x}_1, t_1; \boldsymbol{x}_2, t_2; \ldots; \boldsymbol{x}_n, t_n) = p(\boldsymbol{x}_{n+1}, t_{n+1} | \boldsymbol{x}_n, t_n) ,$$

the stochastic process is known as Markov process. This property, by which the conditional probability at t_{n+1} is uniquely determined by the knowledge of the most recent value X_n, is called Markov property. The conditional probability density $p(\boldsymbol{x}_{n+1}, t_{n+1} | \boldsymbol{x}_n, t_n)$ is called transition density. Using this property, any joint probability density is determined with the only knowledge of the probability density at initial time t_1 and all the transition density functions:

$$\begin{aligned} p(\boldsymbol{x}_1, t_1; \boldsymbol{x}_2, t_2; \ldots; \boldsymbol{x}_n, t_n) &= p(\boldsymbol{x}_n, t_n | \boldsymbol{x}_1, t_1; \boldsymbol{x}_2, t_2; \ldots; \boldsymbol{x}_{n-1}, t_{n-1}) p(\boldsymbol{x}_1, t_1; \boldsymbol{x}_2, t_2; \ldots; \boldsymbol{x}_{n-1}, t_{n-1}) \\ &= p(\boldsymbol{x}_n, t_n | \boldsymbol{x}_{n-1}, t_{n-1}) p(\boldsymbol{x}_1, t_1; \boldsymbol{x}_2, t_2; \ldots; \boldsymbol{x}_{n-1}, t_{n-1}) \\ &= p(\boldsymbol{x}_n, t_n | \boldsymbol{x}_{n-1}, t_{n-1}) p(\boldsymbol{x}_{n-1}, t_{n-1} | \boldsymbol{x}_{n-2}, t_{n-2}) \cdots p(\boldsymbol{x}_1, t_1) . \end{aligned} \tag{2.1}$$

By means of this property, mathematical manipulations with Markov processes often become tractable. For this reason they are widely used for modelling in the natural and social sciences (Gardiner, 2009).

2.1.2 Differential Chapman-Kolmogorov equations

By integrating Equation 2.1 over the set of variables $\boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n-1}$ and then multiplying by the inverse of $p(\boldsymbol{x}_1, t_1)$, we obtain the following relation between the transition probability densities:

$$p(\boldsymbol{x}_n, t_n | \boldsymbol{x}_1, t_1) = \int \cdots \int \mathrm{d}\boldsymbol{x}_2 \cdots \mathrm{d}\boldsymbol{x}_{n-1} \, p(\boldsymbol{x}_n, t_n | \boldsymbol{x}_{n-1}, t_{n-1}) \cdots p(\boldsymbol{x}_2, t_2 | \boldsymbol{x}_1, t_1) .$$

From this, considering only three states of the stochastic variable, $\mathbf{X}(t') = \boldsymbol{y}$, $\mathbf{X}(t'') = \boldsymbol{z}$, $\mathbf{X}(t) = \boldsymbol{x}$, at successive times $t' \leq t'' \leq t$, we obtain the following equation

$$p(\boldsymbol{x}, t | \boldsymbol{y}, t') = \int \mathrm{d}\boldsymbol{z} \, p(\boldsymbol{x}, t | \boldsymbol{z}, t'') \, p(\boldsymbol{z}, t'' | \boldsymbol{y}, t') ,$$

which is known as Chapman-Kolmogorov equation. All conditional probability densities of Markov processes obey the Chapman-Kolmogorov equation, which is simply a consequence of

[1] Here we will use the terms probability density and probability distribution indistinctly.

the Markov property. The meaning of the Chapman-Kolmogorov equation is that the transition density from an initial state $\boldsymbol{y}$ to a final state $\boldsymbol{x}$ is determined by the integral (a sum in a discrete state space) of the transition densities from all intermediary states.

If the following limits of small δt exist for all $\epsilon > 0$

$$\lim_{\delta t \to 0} \Big[\frac{1}{\delta t} \quad p(\boldsymbol{z}, t+\delta t|\boldsymbol{x}, t)\Big] = W(\boldsymbol{z}|\boldsymbol{x}, t)$$

$$\lim_{\delta t \to 0} \Big[\frac{1}{\delta t} \int\limits_{|\boldsymbol{z}-\boldsymbol{x}|<\epsilon} \mathrm{d}\boldsymbol{z}\,(x_i - z_i)\; p(\boldsymbol{z}, t+\delta t|\boldsymbol{x}, t)\Big] = A_i(\boldsymbol{x}, t) + O(\epsilon)$$

$$\lim_{\delta t \to 0} \Big[\frac{1}{\delta t} \int\limits_{|\boldsymbol{z}-\boldsymbol{x}|<\epsilon} \mathrm{d}\boldsymbol{z}(z_i - x_i)(z_j - x_j)\, p(\boldsymbol{z}, t+\delta t|\boldsymbol{x}, t)\Big] = B_{ij}(\boldsymbol{x}, t) + O(\epsilon)$$

then we can write the Chapman-Kolmogorov equation in its differential form[2]

$$\begin{aligned}\frac{\partial}{\partial t}p(\boldsymbol{x}, t|\boldsymbol{y}, t') &= -\sum_i \frac{\partial}{\partial x_i}\Big[A_i(\boldsymbol{x}, t)p(\boldsymbol{x}, t|\boldsymbol{y}, t')\Big] + \frac{1}{2}\sum_{i,j}\frac{\partial^2}{\partial x_i \partial x_j}\Big[B_{ij}(\boldsymbol{x}, t)p(\boldsymbol{x}, t|\boldsymbol{y}, t')\Big] + \\ &\int \mathrm{d}\boldsymbol{z}\Big[W(\boldsymbol{x}|\boldsymbol{z}, t)p(\boldsymbol{z}, t|\boldsymbol{y}, t') - W(\boldsymbol{z}|\boldsymbol{x}, t)p(\boldsymbol{x}, t|\boldsymbol{y}, t')\Big], \end{aligned} \tag{2.2}$$

where A_i and B_{ij} are the functions defined above (Kloeden and Platen, 1992; Gardiner, 2009). The term $W(\boldsymbol{x}|\boldsymbol{z}, t)$ represents the infinitesimal frequency (probability per unit time) with which the process performs jumps greater than ϵ from state $\boldsymbol{z}$ to state $\boldsymbol{x}$. The functions A_i and B_{ij} represent mean and covariance of the infinitesimal difference between final and initial state; they are called drift and diffusion and indicate the instantaneous rate of change of mean and squared fluctuations of the stochastic process, respectively. The matrix with B_{ij} as elements is called diffusion matrix, and by definition is a symmetric positive semidefinite matrix.

The differential Chapman-Kolmogorov equation 2.2 is called forward differential Chapman-Kolmogorov equation, as it describes the evolution of the transition probability density of the stochastic process forward in time. It is possible to derive another differential Chapman-Kolmogorov equation, called backward differential Chapman-Kolmogorov equation, which simply describes the evolution of the transition probability backward in time[3]:

$$\begin{aligned}\frac{\partial}{\partial t'}p(\boldsymbol{x}, t|\boldsymbol{y}, t') &= -\sum_i A_i(\boldsymbol{y}, t')\frac{\partial}{\partial y_i}p(\boldsymbol{x}, t|\boldsymbol{y}, t') - \frac{1}{2}\sum_{i,j}B_{ij}(\boldsymbol{y}, t')\frac{\partial^2}{\partial y_i \partial y_j}p(\boldsymbol{x}, t|\boldsymbol{y}, t') + \\ &\int \mathrm{d}\boldsymbol{z}\, W(\boldsymbol{z}|\boldsymbol{y}, t')\Big[p(\boldsymbol{x}, t|\boldsymbol{y}, t') - p(\boldsymbol{x}, t|\boldsymbol{z}, t')\Big]. \end{aligned} \tag{2.3}$$

Instead of holding fixed the initial conditions ($\boldsymbol{y}$, t'), the backward equation is derived by holding fixed the final conditions ($\boldsymbol{x}$, t) and obtaining the evolution for time $t' \leq t$.

[2]The first limit must be valid uniformly in $\boldsymbol{x}$, $\boldsymbol{z}$ and t, for $|\boldsymbol{x} - \boldsymbol{z}| \geq \epsilon$; the second and third limit must be valid uniformly in $\boldsymbol{z}$, t and ϵ. Note that in Chapman-Kolmogorov equation we are using the functions A_i and B_{ij} defined above with the random variable $\boldsymbol{x}$ as argument. We are considering the time development with respect to final variables $\boldsymbol{x}$, t, given initial variables $\boldsymbol{y}$, t'. The variable $\boldsymbol{z}$ represents an arbitrary state.

[3]We are considering the time development with respect to initial variables $\boldsymbol{y}$, t', given final variables $\boldsymbol{x}$, t. The variable $\boldsymbol{z}$ represents again an intermediate state. Note that the partial derivatives are with respect to the initial variables.

2.1.3 Common Markov processes

From the general Markov process we can distinguish three particular Markov processes: jump process, diffusion process and deterministic process. The first arises when both drift term A_i and diffusion term B_{ij} are zero. The resulting equation is called master equation

$$\frac{\partial}{\partial t}p(\boldsymbol{x},t|\boldsymbol{y},t') = \int \mathrm{d}\boldsymbol{z}\Big[W(\boldsymbol{x}|\boldsymbol{z},t)p(\boldsymbol{z},t|\boldsymbol{y},t') - W(\boldsymbol{z}|\boldsymbol{x},t)p(\boldsymbol{x},t|\boldsymbol{y},t')\Big]$$

and is defined only by jump conditions. A typical sample path of a jump process is discontinuous and represented as a straight line at a given state with jumps to other states occurring with rates defined by the W terms. In presence of only two discrete states x and x', the process becomes a 2-state Markov jump process and is also called *telegraph process*. The master equation simplifies to

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\,p(x\,,t|x_0,t_0) &= -W(x'|x)p(x\,,t|x_0,t_0) + W(x|x')p(x',t|x_0,t_0)\,,\\ \frac{\mathrm{d}}{\mathrm{d}t}\,p(x',t|x_0,t_0) &= -W(x|x')p(x',t|x_0,t_0) + W(x'|x)p(x\,,t|x_0,t_0)\,,\end{aligned}$$

where x_0 is the state at initial time t_0. On the other hand, when the general stochastic process satisfies the following continuity condition (Lindeberg condition)

$$\lim_{\delta t\to 0}\Big[\frac{1}{\delta t}\int\limits_{|\boldsymbol{x}-\boldsymbol{z}|>\epsilon} p(\boldsymbol{x},t+\delta t|\boldsymbol{z},t)\Big] = 0\,,$$

then the jump terms W are zero and the sample paths become continuous. The process in this case is called diffusion process and obeys the Fokker-Planck equation

$$\frac{\partial}{\partial t}p(\boldsymbol{x},t|\boldsymbol{y},t') = -\sum_i \frac{\partial}{\partial x_i}\Big[A_i(\boldsymbol{x},t)p(\boldsymbol{x},t|\boldsymbol{y},t')\Big] + \frac{1}{2}\sum_{i,j}\frac{\partial^2}{\partial x_i\partial x_j}\Big[B_{ij}(\boldsymbol{x},t)p(\boldsymbol{x},t|\boldsymbol{y},t')\Big]\,.$$

The simplest diffusion process is known as Wiener process, commonly denoted with the random variable w[4]. It is described by the following Fokker-Planck equation

$$\frac{\partial}{\partial t}p(w,t|w_0,t_0) = \frac{1}{2}\frac{\partial^2}{\partial w^2}p(w,t|w_0,t_0)\,,$$

where the drift coefficient is 0, the diffusion coefficient is 1 and we assume that the process has value w_0 at time t_0. The solution of this equation is a Gaussian distribution $\mathcal{N}(w|w_0,|t-t_0|)$ with variance increasing in time, so continuous sample paths of the Wiener process can be very different between each other.

Finally, when both jump and diffusion terms are null, the stochastic process reduces to the following equation

$$\frac{\partial}{\partial t}p(\boldsymbol{x},t|\boldsymbol{y},t') = -\sum_i \frac{\partial}{\partial x_i}\Big[A_i(\boldsymbol{x},t)p(\boldsymbol{x},t|\boldsymbol{y},t')\Big]\,,$$

whose solutions, with initial condition $p(\boldsymbol{x},t'|\boldsymbol{y},t') = \delta(x-y)$, give sample paths which satisfy

[4]We will not use W to make a distinction with the jump terms.

an ordinary differential equation (ODE) (Gardiner, 2009).

2.1.4 Stochastic differential equations

Some classes of stochastic processes can be described using stochastic differential equations (SDEs), that are differential equations with random functions of time (with given properties) in the coefficients. They can be conventionally divided in two categories: linear and nonlinear. Linear SDEs can be additive, where random functions are in the inhomogeneous term, or multiplicative, where random functions are in the coefficients that multiply the dependent variable. Nonlinear SDE are nonlinear in the dependent variable and can possibly have further subdivisions.

Here we focus on linear SDEs, whose first historical example was given by Paul Langevin (Langevin, 1906). The Langevin equation has the following form:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = a(x,t) + b(x,t)\xi(t)\,,$$

where $a(x,t) = a_1(t)x(t)+a_2(t)$ and $b(x,t) = b_1(t)x(t)+b_2(t)$ are known linear functions[5]. The term $\xi(t)$ is a rapidly fluctuating function which is mathematically described as a white Gaussian process[6].

In order to solve the equation, we have to compute the integral of the white Gaussian process term. It can be shown that the integral of $\xi(t)$ is interpreted as Wiener process $w(t)$[7], so it follows that $\mathrm{d}w(t) = \xi(t)\mathrm{d}t$ (Gardiner, 2009). By using this relation, the SDE can be meaningfully expressed as the following integral equation

$$x(t) - x(t_0) = \int_{t_0}^{t} a\left[x(s), s\right]\mathrm{d}s + \int_{t_0}^{t} b\left[x(s), s\right]\mathrm{d}w(s)\,,$$

where the first integral is a Riemann integral and the second integral has been interpreted by Itō in a mean square sense[8]. If the quantity $x(t)$ satisfies this integral equation, we say that it obeys the following Itō SDE:

$$\mathrm{d}x(t) = a\left[x(t), t\right]\mathrm{d}t + b\left[x(t), t\right]\mathrm{d}w(t)\,.$$

In order to simulate this linear SDE we can use the Euler-Maruyama method (Iacus, 2008). By choosing a suitable time step Δt, the SDE can be solved in discrete time as

$$\Delta x(k) = x(k+1) - x(k) = a\left[x(k), k\right]\Delta t + b\left[x(k), k\right]\Delta w(k)\,.$$

where $\Delta w(k) \sim \mathcal{N}(0, \Delta t) \sim \sqrt{\Delta t}\cdot\mathcal{N}(0,1)$, since the increments of the Wiener process are inde-

[5] We have indeed generalised the Langevin equation, by defining $b(x,t)$ as a function of the state $x(t)$ and not of the only time t as in the original Langevin equation.

[6] Samples from $\xi(t)$ are normally distributed, with $\langle\xi(t)\rangle = 0$ and $\langle\xi(t)\xi(t')\rangle = \delta(t-t')$.

[7] Mathematically this is a paradox, because continuous sample paths of the Wiener process can be shown to be nowhere differentiable.

[8] Another interpretation in the mean square sense has been given by Stratonovich (Gardiner, 2009).

pendent. Note that given $x(k)$, $x(k+1)$ depends only on the Gaussian independent increments of the Wiener process and is independent of the history of the process up to time k. Therefore the process generated by the SDE is a Gaussian Markov process[9]. It is possible to show (Gardiner, 2009) that the SDE above is associated with a diffusion process with drift coefficient $a(x,t)$ and diffusion coefficient $b(x,t)$ which obeys the following Fokker-Planck equation:

$$\frac{\partial}{\partial t}p(x,t|x_0,t_0) = -\frac{\partial}{\partial x}\Big[a(x,t)p(x,t|x_0,t_0)\Big] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\Big[b(x,t)^2 p(x,t|x_0,t_0)\Big]\,.$$

In addition to the Wiener process, another commonly used Gaussian Markov process is the Ornstein-Uhlenbeck (OU) process, which corresponds to the following SDE:

$$\mathrm{d}x(t) = -kx(t)\mathrm{d}t + \sqrt{D}\mathrm{d}w(t)\,.$$

It can be viewed as an extension of the Wiener process with an additional linear drift term. Solution of this equation can be obtained directly or by solving the associated Fokker-Planck equation

$$\frac{\partial}{\partial t}p(x,t|x_0,t_0) = \frac{\partial}{\partial x}\Big[kxp(x,t|x_0,t_0)\Big] + \frac{1}{2}D\frac{\partial^2}{\partial x^2}p(x,t|x_0,t_0)\,.$$

If we compute the solution, as expected this is a Gaussian distribution with statistics

$$\begin{aligned}\langle x(t)\rangle &= x_0\exp(-kt)\,,\\ \mathrm{var}\,[x(t)] &= \frac{D}{2k}\Big[1-\exp(-2kt)\Big].\end{aligned}$$

Therefore, in contrast to the Wiener process, the OU process admits a stationary distribution for $t\to\infty$. In general we can define a multivariate OU process for a state variable $\boldsymbol{x}(t)\in\mathcal{R}^d$ as

$$\mathrm{d}\boldsymbol{x}(t) = -\boldsymbol{A}\boldsymbol{x}(t)\mathrm{d}t + \sqrt{\boldsymbol{\Sigma}}\mathrm{d}\boldsymbol{w}(t)\,,$$

where the terms $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$ now are $d\times d$ matrices, and $\boldsymbol{w}(t)$ indicates a multivariate Wiener process. The multivariate OU process is completely defined by its first two statistics, the mean and the covariance function, as any other Gaussian stochastic process.

2.1.5 Equations for the moments: univariate linear case

Instead of computing these statistics from the solution of the SDE, it is much simpler to solve equations for the moments which can be directly derived from the SDE. To show this, we consider a single variable linear SDE

$$\mathrm{d}x(t) = \Big[a(t)+b(t)x(t)\Big]\mathrm{d}t + \Big[f(t)+g(t)x(t)\Big]\mathrm{d}w(t)\,,$$

where drift and diffusion terms are given by inhomogeneous linear functions, $a(t)+b(t)x(t)$ and $f(t)+g(t)x(t)$. By Taylor expanding up to the second-order the equation for the general moment

[9] Indeed we obtain a Gaussian transition density only if the drift is linear and the diffusion coefficient is state independent.

$x(t)^n$ we obtain

$$\begin{aligned}
\mathrm{d}\Big[x(t)^n\Big] &= nx(t)^{n-1}\mathrm{d}x(t) + \frac{1}{2}n(n-1)x(t)^{n-2}\Big[\mathrm{d}x(t)\Big]^2 \\
&= nx(t)^{n-1}\mathrm{d}x(t) + \frac{1}{2}n(n-1)x(t)^{n-2}\Big[[a(t)+b(t)x(t)]\mathrm{d}t + [f(t)+g(t)x(t)]\mathrm{d}w(t)\Big]^2 \\
&= nx(t)^{n-1}\mathrm{d}x(t) + \frac{1}{2}n(n-1)x(t)^{n-2}\Big[f(t)+g(t)x(t)\Big]^2\mathrm{d}t\,,
\end{aligned}$$

where we have neglected second-order terms in $\mathrm{d}t$ (and $\mathrm{d}t\,\mathrm{d}w(t)$ terms as well) and we have used the relation $\mathrm{d}w(t)^2 \equiv \mathrm{d}t$ (this can be proved using stochastic calculus (Gardiner, 2009)). Then, replacing $\mathrm{d}x(t)$ and computing the expectation[10], we obtain the equation for the moments:

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle x^n\rangle = \langle x^n\rangle\Big[nb(t) + \frac{1}{2}n(n-1)g(t)^2\Big] + \langle x^{n-1}\rangle\Big[na(t) + n(n-1)f(t)g(t)\Big] + \langle x^{n-2}\rangle\frac{1}{2}n(n-1)f(t)^2\,.$$

The equation for the mean $m = \langle x(t)\rangle$ of the stochastic process is simply obtained by setting $n = 1$:

$$\frac{\mathrm{d}}{\mathrm{d}t}m(t) = b(t)m(t) + a(t)\,.$$

The equation for the variance σ^2 can be computed by the equations for the first and second moment (using the relation $\sigma^2 = \langle x^2\rangle - \langle x\rangle^2$):

$$\frac{\mathrm{d}}{\mathrm{d}t}\sigma(t)^2 = \sigma(t)^2\Big[2b(t) + g(t)^2\Big] + \Big[f(t) + g(t)m(t)\Big]^2\,.$$

2.1.6 Equations for the moments: general case

The previous equations for the moments can be easily derived from a general case. If we have a multivariate stochastic process described by the following nonlinear SDE

$$\mathrm{d}\boldsymbol{x}(t) = \boldsymbol{A}(\boldsymbol{x},t)\mathrm{d}t + \boldsymbol{B}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{w}(t)\,,$$

then, by using Itō's lemma from stochastic calculus (Gardiner, 2009), the evolution of an arbitrary function of $\boldsymbol{x}(t)$, $\boldsymbol{f}(\boldsymbol{x})$, is described by

$$\mathrm{d}\boldsymbol{f}(\boldsymbol{x}) = \left\{\sum_i \boldsymbol{A}_i(\boldsymbol{x},t)\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_i} + \frac{1}{2}\sum_{i,j}\Big[\boldsymbol{B}(\boldsymbol{x},t)\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{x},t)\Big]_{ij}\frac{\partial^2\boldsymbol{f}(\boldsymbol{x})}{\partial x_i\partial x_j}\right\}\mathrm{d}t + \sum_{i,j}\boldsymbol{B}_{ij}(\boldsymbol{x},t)\frac{\partial\boldsymbol{f}(\boldsymbol{x})}{\partial x_i}\mathrm{d}w_j\,.$$

The expectation of this equation becomes

$$\mathrm{d}\,\langle\boldsymbol{f}(\boldsymbol{x})\rangle = \left\langle\sum_i \boldsymbol{A}_i(\boldsymbol{x},t)\frac{\partial}{\partial x_i}\boldsymbol{f}(\boldsymbol{x})\right\rangle\mathrm{d}t + \frac{1}{2}\left\langle\sum_{i,j}\Big[\boldsymbol{B}(\boldsymbol{x},t)\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{x},t)\Big]_{ij}\frac{\partial^2}{\partial x_i\partial x_j}\boldsymbol{f}(\boldsymbol{x})\right\rangle\mathrm{d}t\,, \qquad (2.4)$$

where the term including the Wiener process is zero, since $\langle\mathrm{d}w(t)\rangle = 0$. Equation 2.4 represents a general formula from which we can obtain equations for all the moments. If $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{x}$ we

[10] Recall that from the properties of the Wiener process we have $\langle\mathrm{d}w(t)\rangle = 0$.

easily obtain the equation for the first moment

$$\frac{\mathrm{d}\left\langle \boldsymbol{x} \right\rangle}{\mathrm{d}t} = \left\langle \boldsymbol{A}(\boldsymbol{x}, t) \right\rangle \,,$$

whereas if $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}$ we obtain the equation for the second moment

$$\frac{\mathrm{d}\left\langle \boldsymbol{x}^2 \right\rangle}{\mathrm{d}t} = \left\langle \boldsymbol{x}A^{\mathrm{T}}(\boldsymbol{x}, t) + A(\boldsymbol{x}, t)x^{\mathrm{T}} \right\rangle + \left\langle \boldsymbol{B}(\boldsymbol{x}, t)\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{x}, t) \right\rangle .$$

As before, by using the relationship between first and second moment, we can derive the equation for the covariance matrix $\boldsymbol{P}$ (see Appendix A.1 for derivation):

$$\frac{\mathrm{d}\boldsymbol{P}}{\mathrm{d}t} = \left(\left\langle \boldsymbol{x}A^{\mathrm{T}}(\boldsymbol{x}, t) \right\rangle - \left\langle \boldsymbol{x} \right\rangle \left\langle A^{\mathrm{T}}(\boldsymbol{x}, t) \right\rangle \right) + \left(\left\langle A(\boldsymbol{x}, t)x^{\mathrm{T}} \right\rangle + \left\langle A(\boldsymbol{x}, t) \right\rangle \left\langle x^{\mathrm{T}} \right\rangle \right) + \left\langle \boldsymbol{B}(\boldsymbol{x}, t)\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{x}, t) \right\rangle .$$

From these general equations we can derive the previous equations for the linear univariate case by simply replacing the drift and diffusion terms with $a(t) + b(t)x(t)$ and $f(t) + g(t)x(t)$, respectively. It is important to underline that the general equations for the moments are not simple ODEs. For example the equation for the first moment involves the following expectation

$$\left\langle \boldsymbol{A}(\boldsymbol{x}, t) \right\rangle = \int \boldsymbol{A}(\boldsymbol{x}, t)\, p(\boldsymbol{x}, t) \mathrm{d}\boldsymbol{x} \,,$$

which necessitates the density $p(\boldsymbol{x}, t)$ to be solved, and therefore all the moments of $\boldsymbol{x}(t)$. This is also valid for the equation of the covariance matrix and all the other moments. In contrast, in the linear case, the equations for the mean and the variance depend only on these two statistics. They are uncoupled ODEs which can be solved without requiring the whole probability density.

2.2 Bayesian reasoning

Given noisy observations $\hat{x}$ from a certain dynamical system, we may want to compute a probability distribution which describes that system. This can be done by incorporating the noisy observations in a probabilistic model with some parameters θ representing the properties of the dynamical system. The quantity we are interested in is then a probability distribution $p(\theta|\hat{x})$ over the parameters θ, conditioned on the observations $\hat{x}$. This probability is called posterior probability and is mathematically obtained by application of Bayes' theorem (Robert, 2001):

$$p(\theta|\hat{x}) = \frac{p(\hat{x}|\theta)p(\theta)}{p(\hat{x})} \,,$$

where the distribution $p(\theta)$ over the parameters θ is called prior distribution and $p(\hat{x}|\theta)$ is called likelihood function. The prior distribution represents initial uncertain knowledge about the parameters θ; Bayes' theorem updates this knowledge into the posterior probability, by using information on θ encoded in the observations $\hat{x}$ through the likelihood. The product between prior distribution and likelihood

$$p(\hat{x}|\theta)p(\theta) = p(\theta, \hat{x}) \propto p(\theta|\hat{x})$$

is the joint probability distribution $p(\theta, \hat{x})$, which is proportional to the posterior distribution through a normalising factor

$$p(\hat{x}) = \int p(\hat{x}|\theta)p(\theta)\mathrm{d}\theta\,.$$

This factor is a marginal probability distribution over $\hat{x}$ and is known as evidence or partition function.

Then, in a Bayesian approach both observations and parameters are expressed through probability distributions. The posterior distribution, which is the essence of Bayesian statistics, incorporates the information about the parameters contained both in the prior distribution and in the observations (through the likelihood function). In contrast, a frequentist approach does not associate a probability distribution to the parameters θ, but estimates θ by averaging over many sets of observations $\hat{x}$.

Having a posterior distribution over the parameters means that we can estimate θ by computing not only the mean $\mathbb{E}[\theta|\hat{x}]$ of the distribution but also the uncertainty (e.g. the variance $\sigma^2_{\theta|\hat{x}}$) of this estimation:

$$\begin{aligned}
\mathbb{E}[\theta|\hat{x}] &= \int \theta p(\theta|\hat{x})\mathrm{d}\theta\,,\\
\mathbb{E}[\theta^2|\hat{x}] &= \int \theta^2 p(\theta|\hat{x})\mathrm{d}\theta\,,\\
\sigma^2_{\theta|\hat{x}} &= \mathbb{E}[\theta^2|\hat{x}] - \mathbb{E}[\theta|\hat{x}]^2\,.
\end{aligned}$$

In addition, we might be interested in some function of the parameters $h(\theta)$, whose probabilistic estimate can be computed as

$$\mathbb{E}[h(\theta)|\hat{x}] = \int h(\theta)p(\theta|\hat{x})\mathrm{d}\theta\,,$$

which is another conditional expectation with respect to the posterior distribution $p(\theta|\hat{x})$.

Besides providing a description of the observed process, the Bayesian framework allows a prediction about unseen data (e.g. future observations). This is done by computing the so called predictive distribution,

$$p(x|\hat{x}) = \int p(x|\theta)p(\theta|\hat{x})\mathrm{d}\theta \;= \int p(x,\theta|\hat{x})\mathrm{d}\theta\,,$$

where, in the integral, the first term is the likelihood of the new observation x and the second term is the posterior distribution that was computed at a previous step. This operation of integrating (or summing, in discrete case) to compute a marginal $p(x|\hat{x})$ from a joint distribution $p(x, \theta|\hat{x})$ is called marginalization.

So far we have considered a prior distribution $p(\theta)$ over the parameters θ that we assume to know completely. In a so called fully Bayesian approach, even the prior distribution $p(\theta|\theta')$ might depend on unknown parameters θ' which are called hyperparameters and need to be estimated as well. Hyperparameters can in turn be defined in terms of hyperprior distributions, which can make the marginalization intractable.

One of the drawbacks of Bayesian statistical models is then that the posterior distribution can become very complex when the dimension of the parameter space is large or when

there is not conjugacy property[11]. In these cases, when an explicit form of the posterior distribution cannot be derived, it is possible to resort to approximations which we will present in Section 2.5.

The easier way to estimate parameters θ from noisy observations $\hat{x}$ is by maximising the likelihood $p(\hat{x}|\theta)$. That means to find the set of model parameters θ for which the probability of the observed data $\hat{x}$ is maximised. Although this estimator (known as maximum likelihood) is widely used, it does not take into account the knowledge given through the prior distribution $p(\theta)$. In contrast, another estimation known as maximum a posteriori (MAP), consists in maximising the posterior distribution $p(\theta|\hat{x})$, which means that it finds the parameters θ for which the posterior distribution is maximised. Therefore, MAP incorporates also the information of the prior distribution $p(\theta)$. Indeed, since the marginal distribution $p(\hat{x})$ in the Bayes' theorem does not depend on the parameters θ, the MAP estimator bypasses the computation of $p(\hat{x})$ and is just a maximisation of the joint distribution $p(\theta, \hat{x})$. It is also possible to show that MAP estimation represents a regularised version of the maximum likelihood estimator (Bishop, 2006); the presence of this regularisation (or penalisation) term circumvents the problem of overfitting, from which maximum likelihood estimators suffer.

In a fully Bayesian treatment, where the prior distribution $p(\theta|\theta')$ is parameterised, the estimation of the hyperparameters can be done by choosing the θ' that maximises the marginal likelihood

$$p(\hat{x}|\theta') = \int p(\hat{x}|\theta)p(\theta|\theta')\mathrm{d}\theta = \int p(\hat{x}, \theta|\theta')\mathrm{d}\theta\,.$$

This procedure is called evidence approximation or type II maximum likelihood and it is equivalent to a MAP estimation

$$\begin{aligned} p(\theta'|\hat{x}) &\propto p(\hat{x}|\theta')p(\theta')\,, \\ \theta'_{\mathrm{MAP}} &= \arg\max_{\theta'} p(\theta'|\hat{x})\,, \end{aligned}$$

when the prior over the hyperparameters is weak ($p(\theta') \approx \mathrm{const}$) (Barber, 2012).

Instead of maximising the marginal likelihood, usually it is easier to maximise its logarithm. This can be done in different ways: a first approach consists in computing analytically the evidence $p(\hat{x}|\theta')$ and setting its derivative to zero, in order to obtain equations for θ'. An alternative approach is by means of the expectation-maximisation (EM) algorithm (Dempster et al., 1977), which we will introduce in Section 2.4.

2.3 Graphical model representation

A useful way to describe a probabilistic model is by using a graphical representation. Graphical models (Koller and Friedman, 2009) consist of nodes, representing random variables, and edges, representing the probabilistic relationship between the variables. Two important class of graphical model exist: directed acyclic graphs, also known as Bayesian networks (or belief networks), and undirected graphs, also called Markov random fields. The main difference

[11] The conjugacy property occurs when prior and posterior distribution have the same functional form.

between them is the presence, for directed acyclic graphs, or absence, for undirected graphs, of a direction in the edges. This direction describes a conditional independence relationship between random variables. Other features, as shape and shading of the nodes, make this graphical representation very flexible. Therefore, graphical models have become very popular in machine learning, since they allow an immediate visualisation of the probabilistic model properties. However, it is important to underline that the graphical representation of a probabilistic model is not unique: a probabilistic model can be graphically represented in different ways.

As an example we consider a discrete-time Markov process, commonly known as Markov chain, whose joint probability density is given by

$$p(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n-1}) = p(\boldsymbol{x}_1) \prod_{n=2}^{N} p(\boldsymbol{x}_n|\boldsymbol{x}_{n-1}) \,,$$

where $\boldsymbol{x}_n$ with $n = 1, 2, \ldots, N$ are discrete variables and we have used the Markov property in the last equation. This is the simplest probabilistic model to describe data generated from a dynamical system, where we assume that $\boldsymbol{x}_n$ is uniquely influenced by the variable at the immediate past $\boldsymbol{x}_{n-1}$. Figure 2.1A shows the graphical model representation of a Markov chain: circle nodes, representing discrete-time variables $\boldsymbol{x}_n$, are connected through directed edges which describe the causal relationship between the variables[12]. Here we can view this causal relationship as a temporal ordering, such that the sequence of random variables $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$ constitutes a so called time-series. If we observe a variable $\boldsymbol{x}_n$, which is graphically denoted by a shaded node (Fig. 2.1B), then the joint probability of all the random variables except $\boldsymbol{x}_n$, conditioned on the observation $\boldsymbol{x}_n$, is given by

$$\begin{aligned} p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}, \boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N|\boldsymbol{x}_n) &= \frac{p(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N)}{p(\boldsymbol{x}_n)} \\ &= \frac{p(\boldsymbol{x}_1) \prod_{i=2}^{N} p(\boldsymbol{x}_i|\boldsymbol{x}_{i-1})}{p(\boldsymbol{x}_n)} \\ &= p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}|\boldsymbol{x}_n) p(\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N|\boldsymbol{x}_n) \,, \end{aligned}$$

where we have used Bayes' theorem in the last expression[13]. Then we can say that the set of variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}$ is conditional independent of the set of variables $\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N$ given $\boldsymbol{x}_n$. This can be expressed with the following notation:

$$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1} \perp\!\!\!\perp \boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N \mid \boldsymbol{x}_n \,.$$

The property of conditional independence is fundamental for probabilistic models: it allows to reduce the complexity of the models to obtain a more compact representation (e.g. a factorized distribution), which in turn might facilitate an inference method. A graphical procedure called

[12] By using the term "causal", we are referring just to the temporal ordering of the variables.

[13] We have that $\frac{p(\boldsymbol{x}_1)\cdots p(\boldsymbol{x}_n|\boldsymbol{x}_{n-1})\cdots p(\boldsymbol{x}_N|\boldsymbol{x}_{N-1})}{p(\boldsymbol{x}_n)} = \frac{p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{n-1})p(\boldsymbol{x}_n|\boldsymbol{x}_{n-1})p(\boldsymbol{x}_{n+1},\ldots,\boldsymbol{x}_N|\boldsymbol{x}_n)}{p(\boldsymbol{x}_n)}$. By using the Bayes' rule on the first two terms in the numerator and the term in the denominator we obtain $\frac{p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{n-1})p(\boldsymbol{x}_n|\boldsymbol{x}_{n-1})}{p(\boldsymbol{x}_n)} = \frac{p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{n-1})p(\boldsymbol{x}_n|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{n-1})}{p(\boldsymbol{x}_n)} = p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}|\boldsymbol{x}_n)$.

d-separation (i.e. directed separation (Koller et al., 2007)) is typically used to identify if a set of nodes X in a directed graph is conditionally independent on a second set of nodes Y, given a third set of nodes Z[14]. In the Markov chain's example above, we have implicitly applied one of the rules from the d-separation property and so the variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}$ are said to be d-separated from variables $\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N$ by $\boldsymbol{x}_n$.

An important example of conditional independence is given in Figure 2.1C. Observations $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$ are generated by a multivariate distribution $p(\boldsymbol{x}|\boldsymbol{\eta})$ where $\boldsymbol{\eta}$ represents a parameter of this distribution (e.g. a multivariate mean vector). The joint probability over the observed data

$$p(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N) = \int p(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N|\boldsymbol{\eta}) p(\boldsymbol{\eta}) \mathrm{d}\boldsymbol{\eta}$$

generally cannot factorize[15]. However, by conditioning on $\boldsymbol{\eta}$ (i.e. blocking the paths between the observations), the data become conditionally independent and we can write

$$p(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N|\boldsymbol{\eta}) = \frac{p(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N, \boldsymbol{\eta})}{p(\boldsymbol{\eta})} = \prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{\eta}) \,.$$

In this case, the observations are referred to as independent identically distributed (i.i.d.). This property is useful to describe a set of observations, but it does not allow a representation of the system dynamics. The temporal ordering of the previous Markovian structure is missing in the data and they cannot be considered a proper time-series anymore.

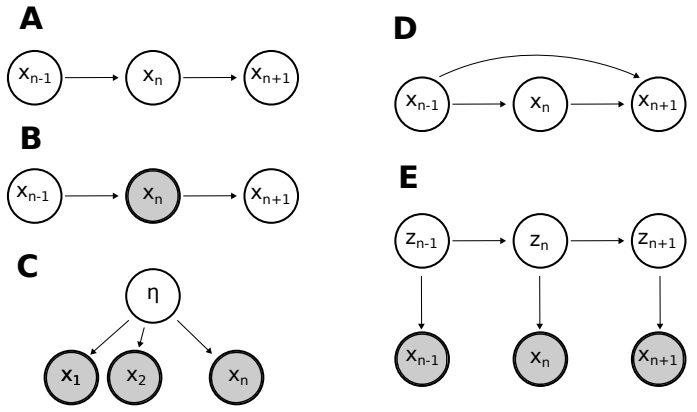

Figure 2.1: Examples of graphical models. A, B: Markov chain; C: i.i.d. condition; D: 2$^{\mathrm{nd}}$ order AR model; E: state space model.

[14]We have that $X \perp\!\!\!\perp Y|Z$ if all the paths between X and Y are blocked, where a path is defined blocked according to the direction of the edges and the status (observed or not observed) of the nodes in the path. Considering the three set of nodes X, Y and Z, four possible situations can be present:

$$X \to Z \to Y \;\;,\;\; X \leftarrow Z \to Y \;\;,\;\; X \leftarrow Z \leftarrow Y \quad \text{and} \quad X \to Z \leftarrow Y \,.$$

The first three cases represent blocked paths if the node in the set Z is observed; in contrast, in the last case, the path is not blocked if the node in the set Z (or one of its descendant) is observed.

[15]Note that we are now considering continuous random variables and so we replaced the summation symbol with an integral.

2.3.1 Latent variables and state space representation

Usually in a real dynamical system the variable at time n is influenced by the trend in the data over the previous variables and not uniquely by the variable at time $n-1$. A way to extend the restrictive Markov chain model in Figure 2.1A is to consider an L^{th} order Markov chain, where the present variable is affected by all the previous L variables (Fig. 2.1D). Then, the joint probability over the variables becomes

$$p(x_1, x_2, \ldots, x_N) = \prod_{n=1}^{N} p(x_n | x_{n-1}, \ldots, x_{n-L}) \,, \qquad \text{with } \; x_n = 0 \; \text{ for } \; n \leq 0 \,.$$

When the variables are continuous and we consider the following linear Gaussian transition densities

$$p(x_n | x_{n-1}, \ldots, x_{n-L}) = \mathcal{N}(x_n | a_1 x_{n-1} + \cdots + a_L x_{n-L}, \sigma^2) \,,$$

where $a_1, \ldots, a_L$ is a set of coefficients and σ^2 the noise variance, then the discrete-time Markov process is known as autoregressive (AR) model (Bishop, 2006). In a more common and compact way it is described by the time law

$$x_n = \sum_{i=1}^{L} a_i x_{n-i} + \epsilon_i \,, \qquad \text{where } \; \epsilon_i = \mathcal{N}(\epsilon_i | 0, \sigma^2) \,.$$

Parameters $a_1, \ldots, a_L$ are called regression coefficients, because they form a linear regression equation to predict the future observation x_n. AR models can be very useful to detect trends in the data, but the number of parameters increases as we use higher order models.

An alternative approach is to introduce latent (or hidden) variables in the model, denoted as $\boldsymbol{z}$ in Figure 2.1E. The result is that the probabilistic model maintains a simple first order Markovian structure in the latent variable space ($\boldsymbol{z}_{n+1} \perp\!\!\!\perp \boldsymbol{z}_{n-1} | \boldsymbol{z}_n$ for each n). However, the conditional independence property is not valid for the observations $\boldsymbol{x}$. In fact, by graphical inspection of the latent variable model, we see that all paths between $\boldsymbol{x}$ nodes are not blocked; therefore the present observation depends (marginally) on all the past observations. Observations become independent only if we condition on the latent variables:

$$\begin{aligned} p(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N | \boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_N) &= \frac{p(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N, \boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_N)}{p(\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_N)} \\ &= \frac{p(\boldsymbol{z}_1) \left[\prod_{n=2}^{N} p(\boldsymbol{z}_n | \boldsymbol{z}_n - 1) \right] \prod_{n=1}^{N} p(\boldsymbol{x}_n | \boldsymbol{z}_n)}{p(\boldsymbol{z}_1) \prod_{n=2}^{N} p(\boldsymbol{z}_n | \boldsymbol{z}_n - 1)} \\ &= \prod_{n=1}^{N} p(\boldsymbol{x}_n | \boldsymbol{z}_n) \,. \end{aligned}$$

Using a latent variable model we then achieve a complexity in the observations' structure which instead is limited by the Markov assumption in the models mentioned above (Bishop, 2006). Another advantage is given by the possibility to model not only variables which are observed,

but also parts of the dynamical system whose observations are not available. This is the case of missing data or model components that cannot be directly measured, but which we are interested in because they are essential to describe the system behaviour.

The latent variable model in Figure 2.1E is also known as state space representation. The most used state space models are hidden Markov models (HMM) (Rabiner, 1989), where latent variables are discrete, and linear dynamical systems (LDS), where both latent and observed variables are continuous and the conditional distributions

$$\begin{aligned} p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1}) &= \mathcal{N}(\boldsymbol{z}_n|\boldsymbol{A}\boldsymbol{z}_{n-1}, \boldsymbol{\Sigma}_w) \qquad \text{with } p(\boldsymbol{z}_1) = \mathcal{N}(\boldsymbol{z}_1|\boldsymbol{m}_0, \boldsymbol{\Sigma}_0) \ , \\ p(\boldsymbol{x}_n|\boldsymbol{z}_n) &= \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{B}\boldsymbol{z}_n, \boldsymbol{\Sigma}_v) \ , \end{aligned}$$

(with $n = 1, 2, \ldots, N$) are linear Gaussian[16]. Here the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are known as transition and emission matrices, while $\boldsymbol{\Sigma}_w$ and $\boldsymbol{\Sigma}_v$ are covariance matrices of noise terms. The vector $\boldsymbol{m}_0$ and the matrix $\boldsymbol{\Sigma}_0$ define the distribution for the initial data. In terms of equations we can write an equation for the latent states $\boldsymbol{z}$ and one for the observations $\boldsymbol{x}$:

$$\boldsymbol{z}_n = \boldsymbol{A}\boldsymbol{z}_{n-1} + \boldsymbol{w}_n \qquad \text{with } \boldsymbol{w} \sim \mathcal{N}(\boldsymbol{w}|0, \boldsymbol{\Sigma}_w) \ , \tag{2.5}$$

$$\boldsymbol{x}_n = \boldsymbol{B}\boldsymbol{z}_n + \boldsymbol{v}_n \qquad \text{with } \boldsymbol{v} \sim \mathcal{N}(\boldsymbol{v}|0, \boldsymbol{\Sigma}_v) \ . \tag{2.6}$$

We can note that a latent variable description introduces an additional source of noise $\boldsymbol{w}$, known as system noise, which is different from the noise in the observations $\boldsymbol{v}$. This is a great advantage to model a dynamical system, because we are assuming that observations $\boldsymbol{x}$ are noisy measurements of latent states $\boldsymbol{z}$. These states, which may represent a real dynamical system, are in turn described by the evolution of a stochastic process. The OU process we mentioned in Section 2.1 can be seen as the continuous-time version of the discrete-time equation for the states' evolution (Eq. 2.5).

2.4 Inference in discretely observed dynamical systems

We consider the problem of estimating the state of a dynamical system from noisy observations. We focus on systems which are described by the Itō SDE

$$\mathrm{d}\boldsymbol{z}(t) = \boldsymbol{f}(\boldsymbol{z}(t))\mathrm{d}t + \boldsymbol{g}(\boldsymbol{z}(t))\mathrm{d}\boldsymbol{w}(t) \, , \tag{2.7}$$

and whose observations are collected at discrete times $n = 1, 2, \ldots, N$:

$$\boldsymbol{x}(n) = \boldsymbol{h}(\boldsymbol{z}(t_n)) + \boldsymbol{v}(n) \, . \tag{2.8}$$

where the n^{th} observation is taken at time t_n. The first equation represents a diffusion process where the vector $\boldsymbol{f}$ and the positive semi-definite matrix $\boldsymbol{g}$ are the drift and diffusion terms, respectively, and where $\boldsymbol{w}(t)$ is a Wiener process. The vector $\boldsymbol{h}$ represents an observation func-

[16] For this reason, LDS are also referred to as linear Gaussian state space models (Barber, 2012).

tion and the measurement noise vector $\boldsymbol{v}(t)$ is distributed according to $\mathcal{N}(0, \boldsymbol{S})$, where $\boldsymbol{S}$ is the observation covariance matrix. This system of equations represents a state space model as the one described in the Section 2.3. Observations are taken at discrete times, but the process $\boldsymbol{z}$ is a continuous-time process which evolves according to the Itō SDE 2.7.

We want to address three different problems, defined as filtering, prediction and smoothing. Given the set of observations $\boldsymbol{x}_n = \boldsymbol{x}(n)$ with $n = 1, 2, \ldots, N$, the filtering problem is computing $p(\boldsymbol{z}(t_N)|\boldsymbol{X}_{1:N})$, that is the estimation of the state $\boldsymbol{z}(t_N)$ given all discrete observations up to time t_N, $\boldsymbol{X}_{1:N} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$. The prediction problem is computing $p(\boldsymbol{z}(t_{>N})|\boldsymbol{X}_{1:N})$, which is the estimation of a future state $z(t_{>N})$, based on past observations. The smoothing problem is finally given by the estimation of $p(\boldsymbol{z}(t_{<N})|\boldsymbol{X}_{1:N})$ (Doucet and Johansen, 2009). The smoothing problem provides a "smoothed" estimation of the unknown sample path $\boldsymbol{z}(t)$, since all observations are used. For this reason, while filtering and prediction can be used for real-time application, the smoothing problem can be only addressed offline.

2.4.1 Evolution of conditional density

We start from the filtering problem: we are interested in finding an estimate of $\boldsymbol{z}(t_N)$ using the information about the observations $\boldsymbol{X}_{1:N} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$. Assuming we are given prior knowledge $p(\boldsymbol{z}_0)$ over the initial data $\boldsymbol{z}(0) = \boldsymbol{z}_0$, the filtering problem consists in computing the posterior density $p(\boldsymbol{z}(t_N)|\boldsymbol{X}_{1:N})$. In other words, we need to find the evolution in time of the conditional density $p(\boldsymbol{z}(t_k)|\boldsymbol{X}_{1:k})$, for all time points t_k. As we mentioned in Section 2.1, the evolution of a diffusion process is given by the Fokker-Planck equation. Here we have a diffusion process which is conditioned on observations at discrete times. Therefore, between observations, it still obeys the Fokker-Planck equation

$$\frac{\partial}{\partial t} p(\boldsymbol{z}(t)|\boldsymbol{X}_{1:k}) = -\sum_i \frac{\partial}{\partial z_i}\Big[f_i(\boldsymbol{z}(t)) p(\boldsymbol{z}(t)|\boldsymbol{X}_{1:k})\Big] + \frac{1}{2}\sum_{i,j} \frac{\partial^2}{\partial z_i \partial z_j}\Big[\boldsymbol{g}_{ij}(\boldsymbol{z}(t)) p(\boldsymbol{z}(t)|\boldsymbol{X}_{1:k})\Big], \quad (2.9)$$

where $t_k \leq t < t_{k+1}$ and the initial condition is $p(\boldsymbol{z}(t)|\boldsymbol{X}_{1:k}) = p(\boldsymbol{z}(t_k)|\boldsymbol{X}_{1:k})$. When an observation $\boldsymbol{x}_k$ is reached, this is incorporated in the conditional density through the likelihood function. This is done by means of Bayes' rule,

$$p(\boldsymbol{z}(t_k)|\boldsymbol{X}_{1:k}) = p(\boldsymbol{z}(t_k)|\boldsymbol{X}_{1:k-1}, \boldsymbol{x}_k) = \frac{1}{Z} p(\boldsymbol{x}_k|\boldsymbol{z}(t_k), \boldsymbol{X}_{1:k-1}) p(\boldsymbol{z}(t_k)|\boldsymbol{X}_{1:k-1})\,. \quad (2.10)$$

By using the conditional independence property, it becomes

$$p(\boldsymbol{z}(t_k)|\boldsymbol{X}_{1:k}) = \frac{1}{Z} p(\boldsymbol{x}_k|\boldsymbol{z}(t_k)) p(\boldsymbol{z}(t_k)|\boldsymbol{X}_{1:k-1})\,, \quad (2.11)$$

where the evidence, $Z = p(\boldsymbol{x}_k|\boldsymbol{X}_{1:k-1})$, is given by the integral of the numerator with respect to $\boldsymbol{z}(t_k)$. Then at observation times, the conditional density $p(\boldsymbol{z}(t_k)|\boldsymbol{X}_{1:k})$ satisfies the following jump condition

$$p(\boldsymbol{z}(t_k^+)|\boldsymbol{X}_{1:k}) = \frac{1}{Z} p(\boldsymbol{x}_k|\boldsymbol{z}(t_k)) p(\boldsymbol{z}(t_k^-)|\boldsymbol{X}_{1:k-1})\,, \quad (2.12)$$

where we have defined

$$p(\boldsymbol{z}(t_k^+)|\boldsymbol{X}_{1:k}) = \lim_{s \to t_k^+} p(\boldsymbol{z}(s)|\boldsymbol{X}_{1:k}) \,, \tag{2.13}$$

$$p(\boldsymbol{z}(t_k^-)|\boldsymbol{X}_{1:k-1}) = \lim_{s \to t_k^-} p(\boldsymbol{z}(s)|\boldsymbol{X}_{1:k-1}) \,, \tag{2.14}$$

with t_k^- and t_k^+ the times before and after the observation $\boldsymbol{x}_k$. Since we assumed a Gaussian observation noise, the likelihood is

$$p(\boldsymbol{x}_k|\boldsymbol{z}(t_k)) = \mathcal{N}(\boldsymbol{x}_k|\boldsymbol{h}(\boldsymbol{z}(t_k)), \boldsymbol{S}) \,. \tag{2.15}$$

The jump condition can then be expressed as

$$p(\boldsymbol{z}(t_k^+)|\boldsymbol{X}_{1:k}) = \frac{\exp\left[-\frac{1}{2}\left(\boldsymbol{x}_k - \boldsymbol{h}_k\right)^{\mathrm{T}} \boldsymbol{S}^{-1} \left(\boldsymbol{x}_k - \boldsymbol{h}_k\right)\right] p(\boldsymbol{z}(t_k^-)|\boldsymbol{X}_{1:k-1})}{\int \exp\left[-\frac{1}{2}\left(\boldsymbol{x}_k - \boldsymbol{h}_k\right)^{\mathrm{T}} \boldsymbol{S}^{-1} \left(\boldsymbol{x}_k - \boldsymbol{h}_k\right)\right] p(\boldsymbol{z}(t_k^-)|\boldsymbol{X}_{1:k-1}) \mathrm{d}\boldsymbol{z}} \,, \tag{2.16}$$

where $\boldsymbol{h}_k = \boldsymbol{h}(\boldsymbol{z}(t_k))$ to keep the notation uncluttered and where we have canceled out the normalising factor $2\pi^{-\frac{D}{2}} |\boldsymbol{S}|^{-\frac{1}{2}}$ (D is the dimension of the state vector). In the limit that observations have infinite variance, which means they do not provide any certain information, the exponentials in the numerator and denominator tend to 1 and the jumps become null (Jazwinski, 1970):

$$p(\boldsymbol{z}(t_k^+)|\boldsymbol{X}_{1:k}) = p(\boldsymbol{z}(t_k^-)|\boldsymbol{X}_{1:k-1}). \tag{2.17}$$

In the prediction problem we want to estimate the future state of the dynamical system using only past observations. The estimate of $p(\boldsymbol{z}(t_{>N})|\boldsymbol{X}_{1:N})$ is provided by solving the Fokker-Planck equation with the filtering distribution $p(\boldsymbol{z}(t_N)|\boldsymbol{X}_{1:N})$ as initial condition.

2.4.2 Point estimation

Once we have the posterior density, we can eventually produce a point estimation $\boldsymbol{z}(t_n)^\star$ for the latent state $\boldsymbol{z}(t_n)$. According to decision theory (Berger, 1985), this can be done by minimising the expected loss:

$$\mathbb{E}[L] = \int L(\boldsymbol{z}(t_n), \boldsymbol{z}(t_n)^\star) p(\boldsymbol{z}(t_n)|\boldsymbol{X}_{1:n}) \mathrm{d}\boldsymbol{z}(t_n) \,, \tag{2.18}$$

where the loss (or cost) function $L(\boldsymbol{z}(t_k), \boldsymbol{z}(t_k)^\star)$ quantifies the loss of taking the estimate $\boldsymbol{z}(t_k)^\star$ of $\boldsymbol{z}(t_k)$. If we utilise the standard quadratic loss function

$$L(\boldsymbol{z}(t_k), \boldsymbol{z}(t_k)^\star) = \left(\boldsymbol{z}(t_k)^\star - \boldsymbol{z}(t_k)\right)^{\mathrm{T}} \left(\boldsymbol{z}(t_k)^\star - \boldsymbol{z}(t_k)\right) \,, \tag{2.19}$$

then it is possible to show (Jazwinski, 1970) that the estimate $\boldsymbol{z}(t_n)^\star$ which minimises the expected loss, is given by the conditional mean $\mathbb{E}[z(t_n)|X_{1:n}]$. Since the loss function is the squared loss, this estimate is known as minimum squared error (or minimum variance) estimate.

It is interesting to note that in a Bayesian approach, we would maximise the joint posterior

distribution

$$p(\boldsymbol{z}(t_1,\ldots,t_n)|\boldsymbol{X}_{1:n}) = \frac{p(\boldsymbol{X}_{1:n}|\boldsymbol{z}(t_1,\ldots,t_n))p(\boldsymbol{z}(t_1,\ldots,t_n))}{p(\boldsymbol{X}_{1:n})}\,, \tag{2.20}$$

or the marginal posterior $p(\boldsymbol{z}(t_n)|\boldsymbol{X}_{1:n})$[17]. Maximising this conditional density means finding the peak (or mode) of the distribution. If this distribution is Gaussian (linear case) then the mode coincides with the mean of the distribution. Therefore, the maximisation of the posterior is equivalent to find the conditional mean of the distribution, which is the minimum variance estimate.

2.4.3 Evolution of moments

Here we derive equations for the evolution of the moments conditioned on observations. We are in particular interested in the evolution of the conditional mean $\langle \boldsymbol{z}(t_k)|\boldsymbol{X}_{1:k}\rangle$, which represents the minimum squared error estimate.

As mentioned above, the probability density $p(\boldsymbol{z}(t)|\boldsymbol{X}_{1:k})$ for the filtering problem, satisfies the Fokker-Planck equation between observations. Therefore, between observations, it will satisfy the equation for the moments of a general diffusion process (Jazwinski, 1970). In Section 2.1 we have already determined general equations for the mean and covariance matrix of a diffusion process:

$$\begin{aligned}
\frac{\mathrm{d}\boldsymbol{m}}{\mathrm{d}t} &= \mathbb{E}\left[\boldsymbol{f}(\boldsymbol{z},t)\right]\,,\\
\frac{\mathrm{d}\boldsymbol{P}}{\mathrm{d}t} &= \mathbb{E}\left[\boldsymbol{z}f^{\mathrm{T}}(\boldsymbol{z},t)\right] - \mathbb{E}\left[\boldsymbol{z}\right]\mathbb{E}\left[f^{\mathrm{T}}(\boldsymbol{z},t)\right] + \mathbb{E}\left[f(\boldsymbol{z},t)z^{\mathrm{T}}\right] - \mathbb{E}\left[f(\boldsymbol{z},t)\right]\mathbb{E}\left[z^{\mathrm{T}}\right] + \mathbb{E}\left[\boldsymbol{g}(\boldsymbol{z},t)\boldsymbol{g}^{\mathrm{T}}(\boldsymbol{z},t)\right]\,,
\end{aligned}$$

which in the linear case are given by simple uncoupled ODEs. At discrete times, the observations are included by means of jump conditions which are obtained using Bayes' theorem as described above. Obviously, in the prediction problem the conditional mean and covariance matrix obey the general equations for the mean and covariance matrix, without satisfying any jump conditions.

2.4.4 Linear case

In order to solve the general nonlinear filtering and prediction problems, some approximations are needed. On the other hand, as we already mentioned, in the case of linear state and observation models the problem becomes much simpler. Assume we have the following model, where the states evolve according to a linear univariate SDE:

$$\mathrm{d}z(t) = \Big[f(t)z(t)\Big]\mathrm{d}t + g(t)\mathrm{d}w(t)\,, \tag{2.21}$$

$$x_n = h(z(t_n)) + v_n\,, \tag{2.22}$$

where f and g depend on the time variable but not on the state z; $w(t)$ is an univariate Wiener process. The observation noise is distributed according to $\mathcal{N}(0,s^2)$ and h represents the linear

[17] In the linear case, the estimate of $z(t_n)$ obtained by maximising $p(\boldsymbol{z}(t_n)|\boldsymbol{X}_{1:n})$ is the same as the one obtained by maximising $p(\boldsymbol{z}(t_1,\ldots,t_n)|\boldsymbol{X}_{1:n})$ (Jazwinski, 1970).

observation function. Then, the conditional density for the filtering problem $p(z(t_k)|X_{1:k})$ is Gaussian and is described only by its mean and variance functions (Jazwinski, 1970). The general equations for the moments reduce to

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}m(t) &= f(t)m(t)\,,\\ \frac{\mathrm{d}}{\mathrm{d}t}\sigma^2(t) &= 2f(t)\sigma^2(t)+g^2(t)\,,\end{aligned}$$

which are valid between observations, $t_k \le t < t_{k+1}$, and can be solved with standard methods. At observation points, the conditional density satisfies the jump condition 2.12, where both likelihood and prior distribution are Gaussian

$$p(x_k|z(t_k)) = \mathcal{N}\left(x_k|h(z(t_k)),s^2\right)\,, \tag{2.23}$$

$$p(z(t_k)|X_{1:k-1}) = \mathcal{N}\left(z(t_k)|m(t_k),\sigma^2(t_k)\right)\,. \tag{2.24}$$

Therefore the posterior $p(z(t_k)|X_{1:k})$ is again Gaussian and its mean and variance can be computed analytically (see Appendix A.2)

$$m(t_k^+) = \frac{m(t_k^-)\,s^2(t_k)+x_k\,\sigma^2(t_k^-)}{s^2(t_k)+\sigma^2(t_k^-)}\,, \tag{2.25}$$

$$\sigma^2(t_k^+) = \frac{s^2(t_k)\,\sigma^2(t_k^-)}{s^2(t_k)+\sigma^2(t_k^-)}\,, \tag{2.26}$$

where $t_k^{\mp}$ refers to the values before and after the jump[18].

2.4.5 Forward-backward algorithm

We now consider the smoothing problem, which means to compute the posterior probability density $p(\boldsymbol{z}(t_k)|\boldsymbol{X}_{1:T})$, with $k < T$. In other words, we use all collected data $\boldsymbol{X}_{1:T}$ to obtain a more accurate inference of past states $\boldsymbol{z}(t_k)$ of the dynamical system.

For brevity here we treat only the linear case, where the computation of the posterior density can be obtained analytically. The conditional density we are interested in is $p(\boldsymbol{z}(t_k)|\boldsymbol{X}_{1:T})$, which in the linear case is a Gaussian distribution as we have described above. By simple mathematical manipulations we obtain an interesting relation for the conditional density $p(\boldsymbol{z}(t_k)|\boldsymbol{X}_{1:T})$:

$$\begin{aligned}p(\boldsymbol{z}(t_k)|\boldsymbol{X}_{1:T}) &\propto p(\boldsymbol{z}(t_k),\boldsymbol{X}_{1:T})\\ &\propto p(\boldsymbol{z}(t_k),\boldsymbol{X}_{1:k},\boldsymbol{X}_{k+1:T}) = p(\boldsymbol{X}_{k+1:T}|\boldsymbol{z}(t_k),\boldsymbol{X}_{1:k})p(\boldsymbol{z}(t_k),\boldsymbol{X}_{1:k})\\ &\propto p(\boldsymbol{X}_{k+1:T}|\boldsymbol{z}(t_k))p(\boldsymbol{z}(t_k),\boldsymbol{X}_{1:k})\,.\end{aligned} \tag{2.27}$$

In the final equation we used the Markov property or, from a graphical point of view, we can see that the latent state $\boldsymbol{z}(t_k)$ d-separates the future observations from the past observations. The posterior density $p(\boldsymbol{z}(t_k)|\boldsymbol{X}_{1:T})$ is then a Gaussian distribution which is proportional to the

[18] For simplicity, mean and variance in 2.25 and 2.26 have been computed using $h(z(t_k)) = z(t_k)$.

product of two terms: $p(\boldsymbol{X}_{k+1:T}|\boldsymbol{z}(t_k))$, which represents the likelihood of future observations, and $p(\boldsymbol{z}(t_k), \boldsymbol{X}_{1:k})$. The latter is in turn proportional to $p(\boldsymbol{z}(t_k)|\boldsymbol{X}_{1:k})$, which is the solution of the filtering problem, through a constant term $p(\boldsymbol{X}_{1:k})$:

$$p(\boldsymbol{z}(t_k)|\boldsymbol{X}_{1:k}) = \frac{p(\boldsymbol{z}(t_k), \boldsymbol{X}_{1:k})}{p(\boldsymbol{X}_{1:k})}\,. \tag{2.28}$$

As we showed previously, the filtering density is computed by solving the Fokker-Planck equation between observations and solving the relation 2.12 when observations are present. On the other hand, the likelihood term $p(\boldsymbol{X}_{k+1:T}|\boldsymbol{z}(t_k))$ between observations satisfies the backward Fokker-Planck equation (Jazwinski, 1970), which we recall that is obtained by holding fixed final conditions (i.e. future observations $\boldsymbol{X}_{k+1:T}$) and computing the derivatives with respect to initial conditions (i.e. latent state $\boldsymbol{z}(t_k)$)[19]. Assume that the system is described by equations

$$\begin{aligned} \mathrm{d}\boldsymbol{z}(t) &= \Big[\boldsymbol{F}(t)\boldsymbol{z}(t)\Big]\mathrm{d}t + \boldsymbol{G}(t)\mathrm{d}\boldsymbol{w}(t)\,, &(2.29)\\ \boldsymbol{x}_n &= \boldsymbol{h}(\boldsymbol{z}(t_n)) + \boldsymbol{v}_n\,, &(2.30)\end{aligned}$$

where now $\boldsymbol{F}$ and $\boldsymbol{G}$ are matrices and $\boldsymbol{v} \sim \mathcal{N}(0, \boldsymbol{S})$. Then the likelihood of future observations (in $t_k \leq t < t_{k+1}$) obeys

$$\frac{\partial}{\partial t}p(\boldsymbol{X}_{k+1:T}|\boldsymbol{z}(t)) = -\sum_i \boldsymbol{f}_i(t)\boldsymbol{z}(t)\frac{\partial}{\partial z_i}p(\boldsymbol{X}_{k+1:T}|\boldsymbol{z}(t)) - \frac{1}{2}\sum_{i,j}\boldsymbol{G}_{ij}(t)\frac{\partial^2}{\partial z_i \partial z_j}p(\boldsymbol{X}_{k+1:T}|\boldsymbol{z}(t))\,,$$

where $\boldsymbol{f}_n(t)$ represents the n^{th} row of $\boldsymbol{F}(t)$. This partial differential equation (PDE) is solved backward in time between observations. The initial condition (at final time T) is given by the likelihood $p(\boldsymbol{x}_T|\boldsymbol{z}(t_T))$, which in the linear case is a Gaussian

$$p(\boldsymbol{x}_T|\boldsymbol{z}(t_T)) \propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{z}(t_T) - \boldsymbol{x}_T\right)^{\mathrm{T}}\boldsymbol{S}^{-1}\left(\boldsymbol{z}(t_T) - \boldsymbol{x}_T\right)\right\}\,, \tag{2.31}$$

where $\boldsymbol{S}^{-1}$ is the inverse covariance matrix (or precision matrix) for the observation noise.

At discrete times, new (past) observations must be added to the likelihood. When an observation $\boldsymbol{x}_k$ is reached, the updated density $p(\boldsymbol{X}_{k:T}|\boldsymbol{z}(t_k))$ is computed using the simple following relation:

$$\begin{aligned} p(\boldsymbol{X}_{k:T}|\boldsymbol{z}(t_k)) &= p(\boldsymbol{x}_k, \boldsymbol{X}_{k+1:T}|\boldsymbol{z}(t_k)) = p(\boldsymbol{X}_{k+1:T}|\boldsymbol{x}_k, \boldsymbol{z}(t_k))p(\boldsymbol{x}_k|\boldsymbol{z}(t_k))\\ &= p(\boldsymbol{X}_{k+1:T}|\boldsymbol{z}(t_k))p(\boldsymbol{x}_k|\boldsymbol{z}(t_k))\,, &(2.32)\end{aligned}$$

where again the last equation is given by the d-separation property. Therefore the new observation is incorporated by using the following jump condition:

$$p(\boldsymbol{X}_{k:T}|\boldsymbol{z}(t_k^-)) = p(\boldsymbol{X}_{k+1:T}|\boldsymbol{z}(t_k^+))p(\boldsymbol{x}_k|\boldsymbol{z}(t_k))\,, \tag{2.33}$$

[19] Indeed, in Section 2.1 we described the backward differential Chapman-Kolmogorov equation. The backward Fokker-Planck equation is simply obtained from the backward differential Chapman-Kolmogorov equation when jump terms are null.

where

$$p(\boldsymbol{X}_{k:T}|\boldsymbol{z}(t_k^-)) = \lim_{s\to t_k^-} p(\boldsymbol{X}_{k:T}|\boldsymbol{z}(s))\,, \tag{2.34}$$

$$p(\boldsymbol{X}_{k+1:T}|\boldsymbol{z}(t_k^+)) = \lim_{s\to t_k^+} p(\boldsymbol{X}_{k+1:T}|\boldsymbol{z}(s))\,. \tag{2.35}$$

Here, t_k^+ and t_k^- are the times before and after the observation $\boldsymbol{x}_k$ (in the backward sense) and $p(\boldsymbol{x}_k|\boldsymbol{z}(t_k))$ is the likelihood of the new observation.

The Gaussian conditional density $p(\boldsymbol{X}_{k+1:T}|\boldsymbol{z}(t_k))$ is completely defined by its mean and variance functions. Therefore we are mainly interested in the evolution of these quantities. Since $p(\boldsymbol{X}_{k+1:T}|\boldsymbol{z}(t))$ obeys the backward Fokker-Planck equation, the equations for the evolution of the moments can be derived as before by the equations of the moments of a general diffusion process. The resulting backward equations for the mean vector and the covariance matrix are

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{m}_b(t) &= \boldsymbol{F}(t)\boldsymbol{m}_b(t)\,,\\
\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{C}_b(t) &= \boldsymbol{F}(t)\boldsymbol{C}_b(t) + \boldsymbol{C}_b(t)\boldsymbol{F}^{\mathrm{T}}(t) - \boldsymbol{G}^2(t)\,,
\end{aligned}$$

and they have to be solved backward in time, from the starting conditions given by the likelihood of the last observation $p(\boldsymbol{x}_T|\boldsymbol{z}(t_T))$.

As we mentioned above, the posterior density which solves the smoothing problem is proportional to the following product:

$$p(\boldsymbol{z}(t_k)|\boldsymbol{X}_{1:T}) \propto p(\boldsymbol{z}(t_k)|\boldsymbol{X}_{1:k})\, p(\boldsymbol{X}_{k+1:T}|\boldsymbol{z}(t_k))\,, \tag{2.36}$$

where both terms on the left hand side of the equation are Gaussian. Therefore the moments of the posterior density $p(\boldsymbol{z}(t_k)|\boldsymbol{X}_{1:T})$ can be computed analytically and result as follows:

$$\boldsymbol{m}(t) = \left(\boldsymbol{C}_f^{-1}(t) + \boldsymbol{C}_b^{-1}(t)\right)^{-1}\left(\boldsymbol{C}_f^{-1}(t)\boldsymbol{m}_f(t) + \boldsymbol{C}_b^{-1}(t)\boldsymbol{m}_b(t)\right)\,, \tag{2.37}$$

$$\boldsymbol{\Sigma}(t) = \left(\boldsymbol{C}_f^{-1}(t) + \boldsymbol{C}_b^{-1}(t)\right)^{-1}\,, \tag{2.38}$$

where $\boldsymbol{m}_f$ and $\boldsymbol{C}_f$ represent the moments for the filtered process $p(\boldsymbol{z}(t_k)|\boldsymbol{X}_{1:k})$. The computation of the backward message is independent of the computation of the filtered process (or forward message), therefore they can be implemented in parallel. The combination of the two messages, forward and backward, is known as forward-backward algorithm.

2.4.6 Parameters learning

In the previous subsections we described how to solve the state inference problem by assuming that model parameters are known. Using prior information over the initial data, we have computed the following posterior distribution over the latent states:

$$p\left(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}\right) = \frac{p\left(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\theta}\right)p\left(\boldsymbol{Z}|\boldsymbol{\theta}\right)}{p\left(\boldsymbol{X}|\boldsymbol{\theta}\right)}\,, \tag{2.39}$$

where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are observed and latent variables, respectively, and $\boldsymbol{\theta}$ represent the adjustable parameters in the joint density $p\left(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}\right)$. In reality, the problem of parameter estimation (or learning) is usually coupled with the state inference one.

For a maximum likelihood estimation of the parameters $\boldsymbol{\theta}$, we need to maximise the evidence $p\left(\boldsymbol{X}|\boldsymbol{\theta}\right)$,

$$p(\boldsymbol{X}|\boldsymbol{\theta}) = \int p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\theta}) p(\boldsymbol{Z}|\boldsymbol{\theta}) \mathrm{d}\boldsymbol{Z} = \int p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \mathrm{d}\boldsymbol{Z} \tag{2.40}$$

but in latent variable models this is not possible. In fact we cannot compute the integral of $p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ with respect to $\boldsymbol{Z}$, because the latent variables $\boldsymbol{Z}$ are unknown.

A way to solve the combination of state inference and parameter learning is to use the so called expectation-maximisation (EM) algorithm (Dempster et al., 1977), which is a general two-step iterative optimisation method where values of the parameters are determined using maximum likelihood. Here we give a brief overview of this popular algorithm.

A possible strategy to compute the integral above, is to marginalise out the latent variables. This is what we do in the E-step of the EM algorithm. In detail, the algorithm is based on the following two iterative stages:

- **E-step**

  We start with an initial estimate of the parameters $\boldsymbol{\theta}^{\mathrm{old}}$ and compute the posterior distribution $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})$. This is used to compute the expectation of the log-likelihood $\log p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$,

  $$\mathrm{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}}) = \int p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}}) \log p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \mathrm{d}\boldsymbol{Z}\,,$$

  where the latent states $\boldsymbol{Z}$ are marginalized out. Therefore, $\mathrm{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}})$ is a function of $\boldsymbol{\theta}$ and of our initial guess $\boldsymbol{\theta}^{\mathrm{old}}$.

- **M-step**

  We maximise the expectation of the log-likelihood $\mathrm{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}})$ with respect to $\boldsymbol{\theta}$,

  $$\boldsymbol{\theta}^{\mathrm{new}} = \arg\max_{\boldsymbol{\theta}} \mathrm{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}})\,,$$

  and we use the estimated parameters $\boldsymbol{\theta}^{\mathrm{new}}$ as a new guess in the E-step.

A mathematical proof that the EM algorithm does maximise the likelihood function can be found in (Bishop, 2006). The EM algorithm can be used to do a MAP estimation, by simply including in the M-step a prior knowledge about the parameters (Barber, 2012):

$$\boldsymbol{\theta}^{\mathrm{new}} = \arg\max_{\boldsymbol{\theta}} \left[\mathrm{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}}) + \log p(\boldsymbol{\theta})\right].$$

At every step, the computation of the posterior distribution $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})$ inside the E-step is done by a forward-backward step as we described above (Ghahramani and Hinton, 1996).

As we mentioned in Section 2.2, we can use the EM algorithm to perform a maximum likelihood estimation of the hyperparameters in a fully Bayesian treatment. This is done by simply treating the model parameters as latent variables, which can be marginalised out.

2.5 Approximate inference methods

As described above, in a probabilistic inference model we are interested in a posterior density over latent variables $\boldsymbol{Z}$ given observations $\boldsymbol{X}$, computed through Bayes' rule:

$$p(\boldsymbol{Z}|\boldsymbol{X}) = \frac{p(\boldsymbol{X},\boldsymbol{Z})}{\int p(\boldsymbol{X},\boldsymbol{Z})\mathrm{d}\boldsymbol{Z}}\,.$$

We might also be interested in some expectations with respect to the posterior density, such as the first and the second moment,

$$\begin{aligned}\langle \boldsymbol{Z}\rangle_{p(\boldsymbol{Z}|\boldsymbol{X})} &= \int \boldsymbol{Z}\, p(\boldsymbol{Z}|\boldsymbol{X})\mathrm{d}\boldsymbol{Z} = \frac{\int \boldsymbol{Z}\, p(\boldsymbol{X},\boldsymbol{Z})\mathrm{d}\boldsymbol{Z}}{\int p(\boldsymbol{X},\boldsymbol{Z})\mathrm{d}\boldsymbol{Z}}\,,\\ \left\langle \boldsymbol{Z}^2\right\rangle_{p(\boldsymbol{Z}|\boldsymbol{X})} &= \int \boldsymbol{Z}^2\, p(\boldsymbol{Z}|\boldsymbol{X})\mathrm{d}\boldsymbol{Z} = \frac{\int \boldsymbol{Z}^2\, p(\boldsymbol{X},\boldsymbol{Z})\mathrm{d}\boldsymbol{Z}}{\int p(\boldsymbol{X},\boldsymbol{Z})\mathrm{d}\boldsymbol{Z}}\,.\end{aligned}$$

In both cases we need to compute integrals (or sums in a discrete state space) over all the space of latent variables, which usually are high dimensional integrals. Furthermore, the form of the posterior density could be very complex, such that an analytical computation of the expectations is infeasible. For these reasons we need a tractable approximation to the posterior density.

Here we briefly introduce a range of tractable approximations to the posterior density, which allow us to compute efficiently the quantities we are interested in. These approximations can be broadly separated into two categories: deterministic techniques, which are based on analytical approximations, and sampling based techniques such as Markov chain Monte Carlo methods (Neal, 1993). We treat deterministic approximations, which do not require large computational power as sampling based techniques and can still generate results comparable with exact solutions. The methods we focus on are variational approximations.

2.5.1 Variational methods

Variational methods are a family of deterministic approximations which are based on bounding properties of the partition function. As Monte Carlo methods, they were developed by statistical physicist and are now becoming very popular in the machine learning community to solve intractable inference problems (Jaakkola, 2001). They essentially consist of two steps: the first is to transform the inference problem into an optimisation problem; the second is to look for approximate solutions to the optimisation problem.

We assume that our target probability density can be defined in terms of a given potential $\psi = -\mathrm{E}(\boldsymbol{X},\boldsymbol{Z})$, where $\mathrm{E}(\boldsymbol{X},\boldsymbol{Z})$ is some energy function (MacKay, 2003):

$$p(\boldsymbol{Z}|\boldsymbol{X}) = \frac{p(\boldsymbol{X},\boldsymbol{Z})}{\mathcal{Z}} = \frac{\exp(\psi)}{\int \exp(\psi)\mathrm{d}\boldsymbol{Z}}\,.$$

The optimisation problem is defined by choosing a so called variational density $q(\boldsymbol{Z})$, which can approximate our target density $p(\boldsymbol{Z}|\boldsymbol{X})$, and an objective function $D(q,p)$ to minimise. In order to obtain tractable computations we choose the relative entropy, also known as Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951):

$$D(q,p) = \mathrm{KL}(q(\boldsymbol{Z})\|p(\boldsymbol{Z}|\boldsymbol{X})) = \int q(\boldsymbol{Z}) \log \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X})} \mathrm{d}\boldsymbol{Z} \,. \tag{2.41}$$

The KL divergence $KL(q\|p)$ satisfies the property that it is always positive and becomes null if and only if $q = p$. Another important property is the asymmetry, that is $KL(q\|p) \neq KL(p\|q)$. As objective function we choose $KL(q\|p)$ and not $KL(p\|q)$ because, as we will see, this produces expectations with respect to q and not to the intractable density p (Opper and Winther, 2001).

By expressing the posterior density in the KL divergence we obtain:

$$\begin{aligned} \mathrm{KL}(q(\boldsymbol{Z})\|p(\boldsymbol{Z}|\boldsymbol{X})) &= \int q(\boldsymbol{Z}) \log \frac{q(\boldsymbol{Z})\mathcal{Z}}{p(\boldsymbol{X},\boldsymbol{Z})} \mathrm{d}\boldsymbol{Z} \\ &= \log \mathcal{Z} + \int q(\boldsymbol{Z}) \log q(\boldsymbol{Z}) \mathrm{d}\boldsymbol{Z} - \int q(\boldsymbol{Z}) \log p(\boldsymbol{X},\boldsymbol{Z}) \mathrm{d}\boldsymbol{Z} \\ &= \log \mathcal{Z} - \mathrm{H}[q(\boldsymbol{Z})] + \langle \mathrm{E}(\boldsymbol{X},\boldsymbol{Z}) \rangle_{q(\boldsymbol{Z})} \,, \end{aligned} \tag{2.42}$$

where in the last equation we have defined the entropy $H[q]$ and we have used the definition of the joint density in terms of potential. The factor $-\log \mathcal{Z}$, where $\mathcal{Z}$ is the partition function, is known in the physics community as free energy. Since the free energy is constant with respect to $\boldsymbol{Z}$, the minimisation of the KL divergence is equivalent to minimise the quantity

$$\mathcal{F}(q) = \langle \mathrm{E}(\boldsymbol{X},\boldsymbol{Z}) \rangle_{q(\boldsymbol{Z})} - \mathrm{H}[q(\boldsymbol{Z})] \,,$$

which is known as variational free energy. In other words, we are looking for a variational density $q(\boldsymbol{Z})$ such that the values assigned to the latent variables $\boldsymbol{Z}$ minimise the energy $\mathrm{E}(\boldsymbol{X},\boldsymbol{Z})$ of the system minus an entropy of the $q(\boldsymbol{Z})$ term. The feasibility of the computations depends on the structure of the probability model $\exp(-\mathrm{E}(\boldsymbol{X},\boldsymbol{Z})) = p(\boldsymbol{X},\boldsymbol{Z})$ and on the choice of $q(\boldsymbol{Z})$ (Jaakkola, 2001). The variational density is chosen within a family of tractable distributions and it is a function of some variational parameters β; therefore the optimisation problem reduces to find a set of values β for which the variational free energy is minimal.

By reversing the last expression for the KL divergence (Eq. 2.42), we can view the optimisation problem as

$$\mathcal{F}(q) = -\log \mathcal{Z} + \mathrm{KL}(q(\boldsymbol{Z})\|p(\boldsymbol{Z}|\boldsymbol{X})) \geq -\log \mathcal{Z} \,. \tag{2.43}$$

The minimum of the variational free energy is obtained when the value of KL divergence is zero, that is if the approximate distribution is exactly the target posterior distribution, $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X})$. From the properties of the KL divergence ($\mathrm{KL} \geq 0$) we get that the variational free energy $\mathcal{F}(q)$ is an upper bound on the exact free energy $-\log \mathcal{Z}$, and it is equivalent to the free energy when the KL divergence is zero.

Considering a parameterised prior density $p(\boldsymbol{Z}|\theta)$, we also may want to learn the parameters

θ. By expressing the posterior density as

$$p(\boldsymbol{Z}|\boldsymbol{X},\theta) = \frac{1}{\mathcal{Z}} p(\boldsymbol{X}|\boldsymbol{Z}) p(\boldsymbol{Z}|\theta)\,, \tag{2.44}$$

the KL divergence becomes

$$\mathrm{KL}(q(\boldsymbol{Z})\|p(\boldsymbol{Z}|\boldsymbol{X},\theta)) = \log \mathcal{Z} + \mathrm{KL}(q(\boldsymbol{Z})\|p(\boldsymbol{Z}|\theta)) - \langle \log p(\boldsymbol{X}|\boldsymbol{Z})\rangle_q = \log \mathcal{Z} + \mathcal{F}(q,\theta)\,, \tag{2.45}$$

where $\langle \log p(\boldsymbol{X}|\boldsymbol{Z})\rangle_q$ is the expectation of the log-likelihood under the variational distribution. Therefore instead of minimising the marginal likelihood as in type II maximum likelihood, we can minimise the variational free energy $\mathcal{F}(q,\theta)$ with respect to the parameters θ. A fully Bayesian variational treatment of the learning problem is also possible (Lappalainen and Miskin, 2000).

So far we have described how to transform the inference problem into an optimisation problem but we have not seen any approximations. In fact from the minimisation of the variational free energy in 2.43, we can still recover the free energy. That occurs when the solution of the optimisation problem is $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X})$. The approximation step is obtained by relaxing the function to optimise or by making some approximating assumption over the terms which are involved in such function, i.e. the variational density $q(\boldsymbol{Z})$.

Mean field approximation

Mean field methods are a class of approximations which originates in statistical physics (Parisi, 1988). The rationale behind the approximation is that in large systems of interacting particles, weak couplings between particles can be neglected. Assuming the particles independent of each other, then the system can be described in term of equations for its mean behaviour.

In practice, mean field methods consist of a set of independence properties which make tractable the statistical inference problem. They especially provide an effective way to relax the optimisation problem expressed through a variational approach. This is essentially obtained by making assumptions about the structure of the variational density q, such that subsequent computations (for minimisation of the variational free energy) become tractable.

The simplest possible structure is where all latent variables are independent. By assuming to have a set of N latent variables, $\boldsymbol{Z} = \{Z_1, Z_2, \ldots, Z_N\}$, the variational density $q(\boldsymbol{Z})$ can be factorized as

$$q(\boldsymbol{Z}) = \prod_{i=1}^{N} q_i(Z_i)\,,$$

which is generally referred to as naive mean field approximation. By using such an independence structure, the evaluation of the objective function in the optimisation problem becomes computationally easier. The objective function is minimized with respect to the variational parameters β_i (with $i = 1, 2, \ldots, N$) of each term q_i of the factorized density. Since q_i are independent of each other, they can be optimised independently.

Notice that by imposing a simple structure on the variational density $q(\boldsymbol{Z})$, we have ap-

proximated the original objective function with a simpler one. While the original was a convex function of the variational density $q(\boldsymbol{Z})$[20], the new objective function does not satisfy this property anymore. As a consequence, different schedules for the optimisation of the variational free energy (e.g. different orders in the optimisation of the q_i and different initialisation of q_i as well), generate different local minima (Jaakkola, 2001). An iterative strategy is then usually adopted. A way to assert the quality of resulting approximations is to compare the values of the variational free energy: according to 2.43, a smaller value provides a better approximation to the free energy.

In some cases, the naive mean field approximation cannot guarantee good results, especially when dependencies between the variables cannot be neglected. Dynamical systems represent an exemplary case, where the continuous temporal structure of the variables cannot be ignored. Then other higher-order approximations can be adopted, where the distribution over the hidden variables is not completely factorized but it maintains a certain dependency structure. This class of approximations is referred to as structured mean field approximations and an application of it is given in Chapter 5.

[20]The second derivative with respect to $q(\boldsymbol{Z})$ of the variational free energy can be shown to be always positive.

Chapter 3

Variational inference in Gaussian-jump processes

We introduce a class of processes, called Gaussian-jump processes, which are central to this work. These processes have been used in systems biology to model the dynamics of biological quantities involved in the mechanism of gene expression (Sanguinetti et al., 2009). In this context, a continuous quantity x (e.g. mRNA concentration) is regulated by another entity (e.g. transcription factor), whose state can be well represented by a discrete on/off variable μ. A knowledge of the dynamics of these regulators is fundamental to the understanding of the underlying gene expression mechanism, but so far a direct experimental measure of μ is impractical due to technical limitations. Therefore, Bayesian inference methods become a useful instrument to reconstruct a posterior probability distribution over μ.

The chapter is divided in 8 sections: in Section 3.1 we define Gaussian-jump processes and in Section 3.2 we introduce the continuous-discrete inference problem (i.e. discrete-time observations of a continuous-time process). In Section 3.3 we describe a variational inference approach for Gaussian-jump processes and in Section 3.4 a conditional approximation to make the variational optimisation problem tractable. In Section 3.5 we consider a multivariate Gaussian-jump process with combinatorial interactions of jump terms and describe a mean field approximation to the inference problem. In Section 3.6 we describe an exact inference method to the inference problem and in Section 3.7 we report some results on a simulated data set. Finally, in Section 3.8 we describe an application of the variational method to the study of *E. coli*'s metabolism.

Gaussian-jump processes were first introduced by Sanguinetti et al. (Sanguinetti et al., 2009). They considered a deterministic limit version of these processes[1] and derived algorithms for exact and variational inference. This approach was extended to more complex models, including a stochastic version (Opper et al., 2010) and combinatorial regulation (Opper and Sanguinetti, 2010). Our contribution in this chapter is given by an optimisation algorithm which speeded up the variational inference approach, in the deterministic limit case. This algorithm enabled a generalisation to a combinatorial regulation case, which was applied to *E. coli*'s data in a paper by Rolfe and colleagues (Rolfe et al., 2012).

[1] As it will be clear later, a deterministic version of the Gaussian-jump process is obtained by considering a zero diffusion constant. Therefore we cannot consider them conditionally Gaussian anymore.

3.1 Gaussian-jump processes

Gaussian-jump processes can be described by the following SDE:

$$\mathrm{d}x(t) = f(x,\mu)\mathrm{d}t + \sigma \mathrm{d}w(t) \qquad \text{with} \quad f(x,\mu) = [A\mu(t) + b - \lambda x(t)] \;, \tag{3.1}$$

where A, b and λ represent constant parameters; $w(t)$ is a Wiener process and σ the diffusion constant. The variable $\mu(t)$ is a random discrete variable which is governed by a Markov jump process. Therefore, the Gaussian-jump process can be seen as an extension of an OU process where the drift includes an additional bias b and a jump term $A\mu$.

The Markov jump process is defined by jump (or switching) rates $f(\mu'|\mu)$ (with $\mu' \neq \mu$), which represent probability per unit time. For a small time increment Δt, $f(\mu'|\mu)\Delta t$ is the probability of the system to switch from state μ to state μ'. This means that the infinitesimal transition density can be mathematically written as

$$p(\mu'|\mu) = \lim_{\Delta t \to 0} \left(\delta_{\mu'\mu} + f(\mu'|\mu)\Delta t\right) \;, \tag{3.2}$$

where $\delta_{\mu'\mu}$ is the Kronecker delta[2].

The previous SDE can be extended to the multivariate case with N Gaussian-jump processes and M Markov jump processes:

$$\mathrm{d}\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x},\boldsymbol{\mu})\mathrm{d}t + \sqrt{\boldsymbol{\Sigma}}\mathrm{d}\boldsymbol{w}(t) \qquad \text{with} \quad \boldsymbol{f}(\boldsymbol{x},\boldsymbol{\mu}) = [\boldsymbol{A}\boldsymbol{\mu}(t) + \boldsymbol{b} - \boldsymbol{\Lambda}\boldsymbol{x}(t)] \;, \tag{3.3}$$

where $\boldsymbol{b}$ is a vector, $\boldsymbol{\Lambda}$ is a $N \times N$ diagonal matrix and $\boldsymbol{A}$ is another $N \times M$ matrix whose element A_{ij} represents the interaction of μ_j with x_i. The diffusion term is now a matrix which for simplicity we choose diagonal, $\sqrt{\boldsymbol{\Sigma}} = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_N)$, and $\boldsymbol{w}(t)$ is a multivariate Wiener process.

To keep the subsequent description easier to follow, we consider the single process case as given by Equation 3.1. An extension to multiple processes with combinatorial interactions will be considered in Section 3.5.

By using the Euler-Maruyama method we can represent the Gaussian-jump process as

$$x_{k+1} = x_k + f\left(x_k, \mu_k\right)\Delta t + \sqrt{\Delta t}\,\sigma\,\epsilon_k \;,$$

where Δt is a small time increment and $\epsilon_k \sim \mathcal{N}(0,1)$. This equation can be used to simulate sample paths from the Gaussian-jump process. Sample paths are nonlinear, since the drift term incorporates both a linear term and a Markov jump process dynamics. We will see that this is necessary to model a complex mechanism such as gene expression, where a linear dynamics could be reductive. However, conditioned on the history of the Markov jump process, the SDE is linear and it is associated with a Gaussian Markov process. Therefore, Gaussian-jump processes fall within the class of conditionally Gaussian processes.

[2] We have simplified the notation by removing the time variable. A more correct notation would be the following: $p(\mu_{t+\Delta t} = \mu'|\mu_t = \mu) \simeq \delta_{\mu_{t+\Delta t}\mu_t} + f(\mu'|\mu)\Delta t$.

3.2 Partly observed Gaussian-jump processes

We consider the following problem. We assume that the Gaussian-jump process $x(t)$ is observed only at N discrete times. Observations y_i are assumed to be i.i.d., therefore y_i is given by $x(t_i)$ plus Gaussian noise with variance σ^2_{obs},

$$y_i = x(t_i) + v_i \qquad \text{with } i = 1, 2, \ldots, N \text{ and } v \sim \mathcal{N}(0, \sigma^2_{\text{obs}}) .$$

Using corrupted data $D = \{y_1, y_2, \ldots, y_N\}$, we are interested in the inference of the latent variables ($\mu(t)$ and $x(t)$) and the estimation of the parameters $\theta \equiv [A, b, \lambda]$. We now focus only on the inference problem, given the parameters; learning of the parameters will be considered later.

The process $x(t)$ is not Markovian, since it depends also on the state of the Markov jump process. But if we consider the joint process $[x(t), \mu(t)]$, then this is Markovian. For a small time increment Δt, the infinitesimal joint transition density is then given by[3]

$$p(x', \mu'|x, \mu) = \lim_{\Delta t \to 0} \left(\frac{1}{\sqrt{2\pi\,\sigma^2 \Delta t}} \exp\left[-\frac{(x' - x - f(x,\mu)\Delta t)^2}{2\sigma^2 \Delta t} \right] \cdot \left(\delta_{\mu'\mu} + f(\mu'|\mu)\Delta t \right) \right) , \qquad (3.4)$$

where the first part in the limit is a Gaussian transition density and the second part is as in Equation 3.2.

Given the Markovian nature of the joint process, we can use the forward-backward algorithm to solve the inference problem, as we have described in Subsection 2.4.5. In this case, we will have to use the forward and backward differential Chapman-Kolmogorov equations instead of the Fokker-Planck equations, since we do have jump terms due to the presence of the Markov jump process. This exact inference solution is expensive from a computational perspective, because it involves the numerical solution of coupled PDEs to find the conditional posterior densities. We will describe it in more detail in Section 3.6. In the following section, we treat a tractable solution to the inference problem, using a variational framework.

3.3 Variational approach

We are interested in the posterior distribution $p(\chi, \nu|D, \theta)$, where χ and ν are continuous-time sample paths (e.g. infinite dimensional objects) of the processes $x(t)$ and $\mu(t)$) over some time interval $[0, T]$. D are the noisy observations and θ are the parameters as defined above. By defining a likelihood function $\mathcal{L}$, the posterior density is given by[4]

$$p(\chi, \nu|D, \theta) = \frac{1}{Z} p(\chi, \nu|\theta)\mathcal{L} , \qquad (3.5)$$

[3] Also in Equation 3.4 we have simplified the notation by removing the time variable. A more correct notation would include: $x_{t+\Delta t} = x'$, $\mu_{t+\Delta t} = \mu'$, $x_t = x$, $\mu_t = \mu$.

[4] The likelihood is a function of the sample paths as well. Here we simplify the notation by omitting this dependency and using $\mathcal{L}$ instead of $\mathcal{L}(\chi, \nu)$.

where Z is the partition function and $p(\chi, \nu|\theta)$ is the prior density over the joint process. Since we have assumed i.i.d. observations we have

$$\mathcal{L} = \prod_{i=1}^{N} p(y_i|x(t_i)) = \frac{1}{\sqrt{2\pi}\sigma_{\text{obs}}} \prod_{i=1}^{N} \exp\left[-\frac{(y_i - x(t_i))^2}{2\sigma_{\text{obs}}^2}\right] . \tag{3.6}$$

In order to have a tractable solution of the inference problem, we transform it in an optimisation problem by using a variational inference approach. By defining a variational density $q(\chi, \nu)$ as an approximation to the real posterior density, the target becomes to minimise the KL divergence $KL[q(\chi, \nu)\|p(\chi, \nu|D, \theta)]$. As we have seen in Section 2.5 (see Eq. 2.45), this is equivalent to minimise the variational free energy

$$\mathcal{F}(q) = KL[q(\chi, \nu)\|p(\chi, \nu|\theta)] - \langle \log \mathcal{L} \rangle_q \tag{3.7}$$

which represents an upper bound on the real free energy $-\log Z$. As variational density we chose a process with a drift $g(x, \mu, t) = [B(t)\mu(t) + d(t) + \alpha(t)x(t)]$, where $B(t)$, $d(t)$ and $\alpha(t)$ are variational parameters and the random variable $\mu(t)$ is described by a Markov jump process with switching rates $g(\mu'|\mu, x, t)$. Then the posterior process is governed by the following SDE[5]:

$$\mathrm{d}x(t) = g(x, \mu, t)\mathrm{d}t + \sigma \mathrm{d}w(t) \qquad \text{with } \; g(x, \mu, t) = [B(t)\mu(t) + d(t) + \alpha(t)x(t)] \; . \tag{3.8}$$

The variational density $q(\chi, \nu)$ is defined by a drift term $g(x, \mu, t)$ and switching rates $g(\mu'|\mu, x, t)$ that are time dependent, since they have to take into account of the non-stationarity of the process over the observation period (Archambeau et al., 2007). Note that the switching rates $g(\mu'|\mu, x, t)$ depend also on the state x (the posterior density is in fact proportional to the joint density of x and μ). Then, for a small time increment Δt, the infinitesimal joint transition density is

$$q(x', \mu'|x, \mu) = \lim_{\Delta t \to 0} \left(\frac{1}{\sqrt{2\pi\,\sigma^2 \Delta t}} \exp\left[-\frac{(x' - x - g(x, \mu, t)\Delta t)^2}{2\sigma^2 \Delta t}\right] \cdot \left(\delta_{\mu'\mu} + g(\mu'|\mu, x, t)\Delta t\right) \right) . \tag{3.9}$$

In order to compute the KL divergence between continuous-time sample paths we follow (Archambeau et al., 2007). By discretising time into small time steps Δt, we consider discrete-time sample paths $X = \{x_k\}_{k=0}^K$ and $V = \{\mu_k\}_{k=0}^K$, where $x_k = x(t_k = k\Delta t)$ and $\mu_k = \mu(t_k = k\Delta t)$, respectively. We compute the KL divergence between the discretised measures in the interval $[0, T]$ (with $T = K\Delta t$),

$$KL[q(X, V)\|p(X, V)] = \int \mathrm{d}X \sum_V q(X, V) \log \frac{q(X, V)}{p(X, V)} \, , \tag{3.10}$$

and then obtain the KL between continuous-time sample paths in the limit of $\Delta t \to 0$. We have omitted the dependence on the parameters θ to keep the notation uncluttered. Since both prior and posterior process are (jointly in $x(t)$ and $\mu(t)$) Markov processes, we can write the previous

[5]Note that this is not a Gaussian-jump process as we defined above, because the switching rates depend on the state x as well.

KL divergence as

$$KL[q(X,V)\|p(X,V)] = \int \cdots \int \mathrm{d}x_0 \cdots \mathrm{d}x_K \sum_{\mu_0} \cdots \sum_{\mu_K} q(x_{0:K}, \mu_{0:K}) \log \frac{q(x_0,\mu_0)\prod_{j=0}^{K-1} q(x_{j+1},\mu_{j+1}|x_j,\mu_j)}{p(x_0,\mu_0)\prod_{j=0}^{K-1} p(x_{j+1},\mu_{j+1}|x_j,\mu_j)} . \tag{3.11}$$

After some computations and using the transition densities in equations 3.4 and 3.9 (see Appendix B.1), we obtain the following KL divergence

$$KL[q(\chi,\nu)\|p(\chi,\nu)] = \int_0^T \mathrm{d}t \int \mathrm{d}x \sum_{\mu} \Upsilon(x,\mu,t) q(x,\mu,t) , \tag{3.12}$$

where we have made explicit the dependence on time of the variational posterior single time marginal density and where we have defined

$$\begin{aligned} \Upsilon(x,\mu,t) &= \frac{1}{2\sigma^2}\left[f(x,\mu) - g(x,\mu,t)\right]^2 \\ &+ \sum_{\mu' \neq \mu} \left[g(\mu'|\mu,x,t) \log \frac{g(\mu'|\mu,x,t)}{f(\mu'|\mu)} + f(\mu'|\mu) - g(\mu'|\mu,x,t) \right] . \end{aligned} \tag{3.13}$$

The inference problem then can be solved by minimising the variational free energy with respect to the parameters of the variational density, that are the time-dependent variational parameters of the posterior drift $g(x,\mu,t)$ and the posterior switching rates $g(\mu'|\mu,x,t)$. For the present form of the KL divergence, this problem is not tractable, because it involves the computation of nontrivial expectations with respect to the posterior density $q(x,\mu,t)$. In order to solve the problem, some approximations are needed. In the next section we will present a conditional approximation, whereas in Chapter 5 we will treat a mean field approximation (which is a conditional approximation as well).

3.4 Conditional approximation

We describe a tractable approximation to compute the variational free energy $\mathcal{F}$ in 3.7. This is a conditional approximation to the posterior marginal density of the variable $\mu(t)$. We start from a general case, when the diffusion constant σ is nonzero, and then consider the deterministic limit of $\sigma \to 0$.

The approximation is obtained by relaxing the form of the switching rates of the posterior Markov jump process and making them independent of x: $g(\mu'|\mu,x,t) = g(\mu'|\mu,t)$. Then we can use the master equation to compute the posterior single-time marginal density $q(\mu,t)$:

$$\frac{\mathrm{d}q(\mu,t)}{\mathrm{d}t} = \sum_{\mu' \neq \mu} \left[-g(\mu'|\mu,t) q(\mu,t) + g(\mu|\mu',t) q(\mu',t) \right] ,$$

where, for normalisation property, we have

$$q(\mu',t) = 1 - \sum_{\mu \neq \mu'} q(\mu,t) .$$

From now we restrict to the case where $\mu(t)$ is a discrete binary variable $\mu(t) = \{1, 0\}$. The master equation simplifies to the following:

$$\frac{\mathrm{d}q(\mu,t)}{\mathrm{d}t} = -g_{1-\mu}(t)q(\mu,t) + g_{\mu}(t)q(1-\mu,t)\,, \tag{3.14}$$

where we have used $g_{1-\mu}(t)$ and $g_{\mu}(t)$ to indicate the switching rates $g(1-\mu|\mu,t)$ and $g(\mu|1-\mu,t)$, respectively. By using the fact that $[q(1-\mu,t)+q(\mu)] = 1$, Equation 3.14 can also be written as

$$\frac{\mathrm{d}q(\mu,t)}{\mathrm{d}t} = -[g_{1-\mu}(t) + g_{\mu}(t)]q(\mu,t) + g_{\mu}(t)\,. \tag{3.15}$$

To compute the KL divergence 3.12, we can compute separately the two terms which come from the two bits in Equation 3.13. The first bit is

$$\begin{aligned}\frac{1}{2\sigma^2}\left[f(x,\mu) - g(x,\mu,t)\right]^2 &= \frac{1}{2\sigma^2}\Big[(\alpha(t)+\lambda)^2 x^2(t) \\ &+ 2(\alpha(t)+\lambda)\left[(d(t)-b)x(t) + (B(t)-A)\mu(t)x(t)\right] \\ &+ (d(t)-b)^2 + (B(t)-A)^2\mu^2(t) + 2(B(t)-A)(d(t)-b)\mu(t)\Big]\,,\end{aligned} \tag{3.16}$$

where we have used the posterior drift as defined in 3.8. In the binary case ($\mu(t) = \{1, 0\}$), we have that $\mu^2(t) = \mu(t)$. By computing the expectation with respect to the variational density we obtain

$$\begin{aligned}\int \mathrm{d}x \sum_{\mu} \frac{1}{2\sigma^2}\left[f(x,\mu) - g(x,\mu,t)\right]^2 q(x,\mu,t) &= \frac{1}{2\sigma^2}\Big[(\alpha(t)+\lambda)^2 M_2(t) \\ &+ 2(\alpha(t)+\lambda)\left[(d(t)-b)M_1(t) + (B(t)-A)R(t)\right] \\ &+ (d(t)-b)^2 + (B(t)-A)^2 q(1,t) \\ &+ 2(B(t)-A)(d(t)-b)q(1,t)\Big]\end{aligned} \tag{3.17}$$

where we have used the fact that

$$\int \mathrm{d}x \sum_{\mu} \mu(t) q(x,\mu,t) = \int \mathrm{d}x \Big[1 \cdot q(x,1,t) + 0 \cdot q(x,0,t)\Big] = q(1,t)\,, \tag{3.18}$$

and we have defined the following moments: $M_1(t) = \mathbb{E}_q[x(t)]$, $M_2(t) = \mathbb{E}_q[x^2(t)]$, $R(t) = \mathbb{E}_q[x(t)\mu(t)]$. These moments are found to solve a set of uncoupled ODEs which can be derived directly from the forward differential Chapman-Kolmogorov equation (see Appendix B.2):

$$\frac{\mathrm{d}M_1}{\mathrm{d}t} = \alpha(t)M_1(t) + B(t)q(1,t) + d(t) \tag{3.19}$$

$$\frac{\mathrm{d}M_2}{\mathrm{d}t} = 2\alpha(t)M_2(t) + 2B(t)R(t) + 2d(t)M_1(t) + \sigma^2 \tag{3.20}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = [\alpha(t) - g(t)]R(t) + g_{+}(t)M_1(t) + [B(t) + d(t)]q(1,t)\,. \tag{3.21}$$

We have used $g_{+}(t)$ and $g_{-}(t)$ to refer to switching rates $g(1|0,t)$ and $g(0|1,t)$, respectively, and the variable $g(t)$ for the sum $g(t) = g_{+}(t) + g_{-}(t)$.

The other term in the KL divergence 3.12, comes from the second bit in Equation 3.13.

With the conditional approximation, the posterior switching rates are independent of x, then the second term in 3.13 is constant with respect to x.

We can write the final form of the KL divergence 3.12 in the conditional approximation as

$$\begin{aligned} KL[q(\chi,\nu)\|p(\chi,\nu)] &= \int_0^T \frac{1}{2\sigma^2}\Big[(\alpha(t)+\lambda)^2 M_2(t) + 2(\alpha(t)+\lambda)\left[(d(t)-b)M_1(t) + (B(t)-A)R(t)\right] \\ &+ (d(t)-b)^2 + (B(t)-A)^2 q(1,t) + 2(B(t)-A)(d(t)-b)q(1,t)\Big]\mathrm{d}t \qquad (3.22) \\ &+ \int_0^T \sum_\mu q(\mu,t) \sum_{\mu'\neq\mu} \left[g(\mu'|\mu,t)\log\frac{g(\mu'|\mu,t)}{f(\mu'|\mu)} + f(\mu'|\mu) - g(\mu'|\mu,t)\right]\mathrm{d}t\,, \end{aligned}$$

where $q(\mu,t)$ in the last line, comes from the integration of $q(x,\mu,t)$ in $\mathrm{d}x$.

In this way, the minimisation of the variational free energy $\mathcal{F}$ (Eq. 3.7) becomes a constrained optimisation problem: we need to minimise $\mathcal{F}$, which is a functional of several variables (variational parameters and posterior density), subject to some constraints. These constraints are represented by the equations for the moments 3.19-3.21. In addition, a further constraint is given by the master equation (Eq. 3.14), which links the posterior switching rates to the posterior density. Therefore, minimisation of the functional $\mathcal{F}$ can be done by using calculus of variations (Bishop, 2006), where the constraints can be incorporated using Lagrange multipliers. By introducing Lagrange multipliers $\xi(t)$, $\phi(t)$, $\kappa(t)$ and $\psi(t)$, we obtain the new functional[6]:

$$\begin{aligned} \mathcal{L} &= KL[q(\chi,\nu)\|p(\chi,\nu)] - \langle\log\mathcal{L}\rangle_q \\ &+ \int_0^T \mathrm{d}t\,\xi(t)\left[\frac{\mathrm{d}q(1,t)}{\mathrm{d}t} + g(t)q(1,t) - g_+(t)\right] \\ &+ \int_0^T \mathrm{d}t\,\phi(t)\left[\frac{\mathrm{d}M_1(t)}{\mathrm{d}t} - \alpha(t)M_1(t) - B(t)q(1,t) - d(t)\right] \\ &+ \int_0^T \mathrm{d}t\,\kappa(t)\left[\frac{\mathrm{d}M_2(t)}{\mathrm{d}t} - 2\alpha(t)M_2(t) - 2B(t)R(t) - 2d(t)M_1(t) - \sigma^2\right] \\ &+ \int_0^T \mathrm{d}t\,\psi(t)\left[\frac{\mathrm{d}R(t)}{\mathrm{d}t} - [\alpha(t)-g(t)]R(t) - g_+(t)M_1(t) - [B(t)+d(t)]q(1,t)\right], \qquad (3.23) \end{aligned}$$

where, in the second line, we have included the master equation. The expectation of the log-likelihood (Eq. 3.6) is given by

$$\begin{aligned} \langle\log\mathcal{L}\rangle_q &= \left\langle \log\left[\frac{1}{\sqrt{2\pi}\sigma_{\mathrm{obs}}}\right] - \frac{1}{2\sigma^2_{\mathrm{obs}}}\sum_{i=1}^N [y_i - x(t_i)]^2 \right\rangle_q \\ &= \log\left[\frac{1}{\sqrt{2\pi}\sigma_{\mathrm{obs}}}\right] - \frac{1}{2\sigma^2_{\mathrm{obs}}}\sum_{i=1}^N \left[y_i^2 - 2y_i M_1(t_i) + M_2(t_i)\right], \qquad (3.24) \end{aligned}$$

and it depends on the moments $M_1(t)$ and $M_2(t)$.

[6]Since the constraints must be valid at all time points in the interval $[0,T]$, Lagrange multipliers are continuous functions of time.

3.4.1 Deterministic limit

As mentioned above, we restrict our interest to the deterministic case. In the deterministic limit the diffusion constant σ tends to zero. This means that we must have $g(x, \mu, t) = f(x, \mu)$, because a different choice would lead to an infinite value of the term 3.17 and consequently of the KL divergence. By enforcing $g(x, \mu, t) = f(x, \mu)$ we then obtain the following values for the variational parameters: $B(t) = A$, $d(t) = b$, $\alpha(t) = -\lambda$. The KL divergence 3.22 takes the simple form

$$KL[q(\chi, \nu) \| p(\chi, \nu)] = \int_0^T \mathrm{d}t \sum_{\mu} q(\mu, t) \sum_{\mu' \neq \mu} \left[g(\mu'|\mu, t) \log \frac{g(\mu'|\mu, t)}{f(\mu'|\mu)} + f(\mu'|\mu) - g(\mu'|\mu, t) \right] . \quad (3.25)$$

The Lagrangian becomes

$$\begin{aligned}
\mathcal{L} &= KL[q(\chi, \nu) \| p(\chi, \nu)] - \langle \log \mathcal{L} \rangle_q \\
&+ \int_0^T \mathrm{d}t\, \xi(t) \left[\frac{\mathrm{d}q(1,t)}{\mathrm{d}t} + g(t) q(1,t) - g_+(t) \right] \\
&+ \int_0^T \mathrm{d}t\, \phi(t) \left[\frac{\mathrm{d}M_1(t)}{\mathrm{d}t} + \lambda M_1(t) - A q(1,t) - b \right] \\
&+ \int_0^T \mathrm{d}t\, \kappa(t) \left[\frac{\mathrm{d}M_2(t)}{\mathrm{d}t} + 2\lambda M_2(t) - 2AR(t) - 2bM_1(t) \right] \\
&+ \int_0^T \mathrm{d}t\, \psi(t) \left[\frac{\mathrm{d}R(t)}{\mathrm{d}t} + [\lambda + g(t)] R(t) - g_+(t) M_1(t) - [A + b] q(1,t) \right] , \qquad (3.26)
\end{aligned}$$

where the KL divergence is now given by Equation 3.25 and we have set $\sigma = 0$ in the constraint for the second moment $M_2(t)$. We have included the constraints for the moments, because the likelihood depends on the first two moments. Therefore, we have also included the constraint for $R(t)$, because the second moment $M_2(t)$ depends on it.

Optimisation algorithm

We need to minimise the functional 3.26 with respect to the posterior density and the switching rates (linked through the master equation), with additional constraints represented by equations for the moments $M_1(t)$, $M_2(t)$, $R(t)$. Then, minimisation of 3.26 means to find the stationary points of the Lagrangian with respect to $M_1(t)$, $M_2(t)$, $R(t)$ and $q(1,t)$[7].

The algorithm is the following. We start from an initial guess of the posterior switching rates $g_\pm(t)$ and we solve the master equation and ODEs for the moments. Then we compute

[7] Stationary points of the functional $\mathcal{L}(\boldsymbol{f})$ are the functions $\boldsymbol{f}$ for which $\mathcal{L}(\boldsymbol{f})$ is insensitive to small variations of $\boldsymbol{f}$. This happens if $\boldsymbol{f}$ satisfy the so called Euler-Lagrange equations (Bishop, 2006), which means that the functional derivatives $\delta\mathcal{L}/\delta\boldsymbol{f}$ must vanish.

functional derivatives with respect to the moments $M_1(t)$, $M_2(t)$, $R(t)$ and $q(1,t)$:

$$\frac{\delta\mathcal{L}}{\delta M_1(t)} = -\frac{1}{\sigma^2_{\text{obs}}}\sum_{i=1}^{N} y_i\delta(t-t_i) - \frac{\mathrm{d}\phi(t)}{\mathrm{d}t} + \lambda\phi(t) - 2b\kappa - g_+(t)\psi(t)\,, \tag{3.27}$$

$$\frac{\delta\mathcal{L}}{\delta M_2(t)} = \frac{1}{2\sigma^2_{\text{obs}}}\sum_{i=1}^{N} \delta(t-t_i) - \frac{\mathrm{d}\kappa(t)}{\mathrm{d}t} + 2\lambda\kappa(t)\,, \tag{3.28}$$

$$\frac{\delta\mathcal{L}}{\delta R(t)} = -2A\kappa(t) - \frac{\mathrm{d}\psi(t)}{\mathrm{d}t} + (\lambda + g(t))\psi(t)\,, \tag{3.29}$$

$$\begin{aligned}\frac{\delta\mathcal{L}}{\delta q(1,t)} &= -\frac{\mathrm{d}\xi(t)}{\mathrm{d}t} + g(t)\xi(t) - A\phi(t) - (A+b)\psi(t) \\ &+ \left[g_-(t)\log\frac{g_-(t)}{f_-} + f_- - g_-(t)\right] - \left[g_+(t)\log\frac{g_+(t)}{f_+} + f_+ - g_+(t)\right]\,,\end{aligned} \tag{3.30}$$

where we have used integration by parts and boundary conditions[8]. We have used f_+ and f_- to refer to switching rates $f(1|0)$ and $f(0|1)$. In order for the functional to be stationary with respect to $M_1(t)$, $M_2(t)$, $R(t)$ and $q(1,t)$, the functional derivatives are set to zero. By doing so, we obtain ODEs for the Lagrange multipliers, which must be solved in the following order

$$\frac{\mathrm{d}\kappa(t)}{\mathrm{d}t} = \frac{1}{2\sigma^2_{\text{obs}}}\sum_{i=1}^{N} \delta(t-t_i) + 2\lambda\kappa(t)\,, \tag{3.31}$$

$$\frac{\mathrm{d}\psi(t)}{\mathrm{d}t} = -2A\kappa(t) + (\lambda + g(t))\psi(t)\,, \tag{3.32}$$

$$\frac{\mathrm{d}\phi(t)}{\mathrm{d}t} = -\frac{1}{\sigma^2_{\text{obs}}}\sum_{i=1}^{N} y_i\delta(t-t_i) + \lambda\phi(t) - 2b\kappa - g_+(t)\psi(t)\,, \tag{3.33}$$

$$\begin{aligned}\frac{\mathrm{d}\xi(t)}{\mathrm{d}t} &= g(t)\xi(t) - A\phi(t) - (A+b)\psi(t) \\ &+ \left[g_-(t)\log\frac{g_-(t)}{f_-} + f_- - g_-(t)\right] - \left[g_+(t)\log\frac{g_+(t)}{f_+} + f_+ - g_+(t)\right]\,,\end{aligned} \tag{3.34}$$

backward in time, starting from the final conditions $\kappa(T) = \psi(T) = \phi(T) = \xi(T) = 0$. Finally, by using Lagrange multipliers, moments and $q(1,t)$, we compute the gradients with respect to the functions of interest

$$\frac{\delta\mathcal{L}}{\delta g_+(t)} = q(0,t)\log\frac{g_+(t)}{f_+} - \xi(t)(1-q(1,t)) + (R(t) - M_1(t))\psi(t)\,, \tag{3.35}$$

$$\frac{\delta\mathcal{L}}{\delta g_-(t)} = q(1,t)\log\frac{g_-(t)}{f_-} + q(1,t)\xi(t) + R(t)\psi(t)\,, \tag{3.36}$$

where $q(0,t) = 1 - q(1,t)$. These gradients are used in a gradient descent to update the posterior switching rates. These are in turn used to update the process, through the master equation. The algorithm is iterated until a minimum of the variational free energy is reached. In practice we do not know when the minimum is reached, so we define a threshold and check when the change in variational free energy $|\Delta\mathcal{F}| = |\mathcal{F}_{k+1} - \mathcal{F}_k|$ is below that threshold.

[8]We used integration by parts as follows: $\int_0^T \phi(t)\frac{\mathrm{d}M_1(t)}{\mathrm{d}t}\mathrm{d}t = \left[\phi(t)M_1(t)\Big|_0^T - \int_0^T \frac{\mathrm{d}\phi(t)}{\mathrm{d}t}M_1(t)\mathrm{d}t\right]$. Perturbation is null at final time $t = T$, so $\phi(T) = 0$. For simplicity we also set the value of the moments (and $q(1,t)$) at initial time $t = 0$, thus we do not need to optimise them.

Parameter learning

So far we focused on the inference problem and we considered known parameters $\theta \equiv [A, b, \lambda]$. In reality, the variational free energy $\mathcal{F}$ depends on the parameters θ as well. As we described in Section 2.5, parameter learning can be performed in a variational approach by minimising $\mathcal{F}$ with respect to θ. This can be viewed as an approximation to the type II maximum likelihood.

Minimisation of $\mathcal{F}$ with respect to θ is done by computing the derivatives of the Lagrangian 3.26 with respect to A, b and λ:

$$\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}A} = \int_0^T \mathrm{d}t\Big[-q(1,t)[\phi(t)+\psi(t)] - 2R(t)\kappa(t)\Big], \tag{3.37}$$

$$\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}b} = \int_0^T \mathrm{d}t\Big[-\phi(t) - 2M_1(t)\kappa(t) - q(1,t)\psi(t)\Big], \tag{3.38}$$

$$\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\lambda} = \int_0^T \mathrm{d}t\Big[M_1(t)\phi(t) + 2M_2(t)\kappa(t) + R(t)\psi(t)\Big]. \tag{3.39}$$

The whole optimisation procedure consists in the following gradient descent algorithm: we use an initial guess for the parameters $\theta \equiv [A, b, \lambda]$ to compute the gradients with respect to the switching rates and with respect to the parameters. The gradients are used to update the posterior density and the parameters. By using the updated quantities we compute again the gradients and so on, until a minimum is reached.

3.5 Combinatorial interactions

We now consider the multivariate case where we have N Gaussian-jump processes $\boldsymbol{x}(t) = [x_1(t), x_2(t), \ldots, x_N(t)]$. If each of them is driven by M independent Markov jump processes, the system can be described by Equation 3.3. Then a single Gaussian-jump process $x_k(t)$ obeys the following SDE

$$\mathrm{d}x_k = [A_{k1}\mu_1(t) + \cdots + A_{kM}\mu_M(t) + b_k - \lambda_k x_k(t)]\,\mathrm{d}t + \sigma_k \mathrm{d}w_k(t)\,. \tag{3.40}$$

If we consider the case where each of the Gaussian-jump process is also driven by the interaction of multiple Markov jump processes, the process $x_k(t)$ then obeys the following SDE

$$\mathrm{d}x_k = \left[A_{k1}\mu_1(t) + \cdots + A_{kM}\mu_M(t) + \sum_{i=1}^{M}\left(\sum_{j=i+1}^{M} C_{k\,ij}\mu_i(t)\mu_j(t) \right) + b_k - \lambda_k x_k(t) \right] \mathrm{d}t + \sigma_k \mathrm{d}w_k(t)\,,$$

where we have considered only combinations of couples of Markov jump processes. The additional parameters $C_{k\,ij}$ indicate the combinatorial interaction between $\mu_i(t)$ and $\mu_j(t)$ with $x_k(t)$. The total number of parameters C is given by

$$N\binom{M}{2} = N\frac{M!}{2!(M-2)!}\,. \tag{3.41}$$

Here, we extend the variational approach described in Section 3.4 to this multivariate combinatorial system. For simplicity, we consider a system with N Gaussian-jump processes driven by

only two telegraph processes, in the deterministic limit of $\sigma \to 0$. We can represent this system in the following compact way

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{\mu}(t) + \boldsymbol{C}\mu_1(t)\mu_2(t) + \boldsymbol{b} - \boldsymbol{\Lambda}\boldsymbol{x}(t)\,, \tag{3.42}$$

where $\boldsymbol{A}$ is a $N \times 2$ matrix, $\boldsymbol{\mu}(t) = [\mu_1(t)\ \ \mu_2(t)]^{\mathrm{T}}$ is the vector of telegraph processes, $\boldsymbol{C} = [C_1\ \cdots\ C_N]^{\mathrm{T}}$ is a N-dimensional vector, $\boldsymbol{b} = [b_1\ \cdots\ b_N]^{\mathrm{T}}$ is another N-dimensional vector, $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$ is a $N \times N$ diagonal matrix.

When $\sigma \to 0$, a knowledge of the telegraph processes ($\mu_1(t)$ and $\mu_2(t)$) and of the parameters $\Theta = [\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{C}, \boldsymbol{\Lambda}]$, determines completely the process $\boldsymbol{x}(t)$. Given the parameters, we are interested in the posterior distribution

$$p(\nu_1, \nu_2|D, \Theta) \propto p(D|\nu_1, \nu_2, \Theta)p(\nu_1)p(\nu_2) \tag{3.43}$$

where D are noisy observations of $\boldsymbol{x}(t)$; ν_1 and ν_2 represent continuous-time sample paths of the telegraph processes. Also in this case, we adopt a conditional approximation (posterior switching rates independent of x). The variational free energy to minimise is given by

$$\mathcal{F}(q) = KL[q(\nu_1, \nu_2)\|p(\nu_1, \nu_2)] - \langle \log \mathcal{L} \rangle_q\,. \tag{3.44}$$

As before the KL divergence can be computed between discretised paths $V_1 = \{\mu_{1\,k}\}_{k=0}^{K}$ and $V_2 = \{\mu_{2\,k}\}_{k=0}^{K}$, in the interval $[0, T]$ (with $T = K\Delta t$),

$$KL[q(\nu_1, \nu_2)\|p(\nu_1, \nu_2)] = \lim_{\Delta t \to 0} \sum_{V_1}\sum_{V_2} q(V_1, V_2) \log \frac{q(V_1, V_2)}{p(V_1)p(V_2)}\,, \tag{3.45}$$

which, assuming a factorisation of the posterior process $q(V_1, V_2) = q(V_1)q(V_2)$, becomes

$$\begin{aligned} KL[q(\nu_1, \nu_2)\|p(\nu_1, \nu_2)] &= \lim_{\Delta t \to 0} \left[\sum_{V_1} q(V_1) \log \frac{q(V_1)}{p(V_1)} + \sum_{V_2} q(V_2) \log \frac{q(V_2)}{p(V_2)} \right] \\ &= KL[q(\nu_1)\|p(\nu_1)] + KL[q(\nu_2)\|p(\nu_2)]\,. \end{aligned} \tag{3.46}$$

Note that by using the factorisation for the posterior process, we are considering a mean field type solution to the variational problem (Opper and Sanguinetti, 2010). The two KL divergence terms are computed as in Equation 3.25:

$$\begin{aligned} KL[q(\nu_1)\|p(\nu_1)] &= \int_0^T \mathrm{d}t \sum_{\mu_1} q(\mu_1, t) \sum_{\mu_1' \neq \mu_1} \left[g(\mu_1'|\mu_1, t) \log \frac{g(\mu_1'|\mu_1, t)}{f(\mu_1'|\mu_1)} + f(\mu_1'|\mu_1) - g(\mu_1'|\mu_1, t) \right], \\ KL[q(\nu_2)\|p(\nu_2)] &= \int_0^T \mathrm{d}t \sum_{\mu_2} q(\mu_2, t) \sum_{\mu_2' \neq \mu_2} \left[g(\mu_2'|\mu_2, t) \log \frac{g(\mu_2'|\mu_2, t)}{f(\mu_2'|\mu_2)} + f(\mu_2'|\mu_2) - g(\mu_2'|\mu_2, t) \right]. \end{aligned}$$

The log-likelihood term in the variational free energy is now given by

$$\langle \log \mathcal{L} \rangle_q = \sum_{i=1}^{N} \log \left[\frac{1}{\sqrt{2\pi}\sigma_{i\,\mathrm{obs}}} \right] - \sum_{i=1}^{N}\sum_{j=1}^{D} \frac{1}{2\sigma_{i\,\mathrm{obs}}^2} \left[y_{ij}^2 - 2y_{ij}M_{1i}(t_j) + M_{2i}(t_j) \right], \tag{3.47}$$

where the index j runs over the number of observation points and the index i over the number of processes $x_i(t)$. As before, the variational free energy depends on the first and second moment of the posterior density, $M_{1k}(t) = \mathbb{E}_q[x_k(t)]$ and $M_{2k}(t) = \mathbb{E}_q[x_k^2(t)]$. For the k$^{\text{th}}$ process $x_k(t)$, it is possible to show (see Appendix B.3) that the first and second moment, $M_{1k}(t)$ and $M_{2k}(t)$, obey the following ODEs:

$$\frac{\mathrm{d}M_{1k}(t)}{\mathrm{d}t} = -\lambda_k M_{1k}(t) + A_{k1}q_1(1,t) + A_{k2}q_2(1,t) + C_k q_1(1,t)q_2(1,t) + b_k\,, \tag{3.48}$$

$$\frac{\mathrm{d}M_{2k}(t)}{\mathrm{d}t} = -2\lambda_k M_{2k}(t) + 2A_{k1}R_{1k}(t) + 2A_{k2}R_{2k}(t) + 2C_k R_{12k}(t) + 2b_k M_{1k}(t)\,, \tag{3.49}$$

where we have defined $q_1(1,t) = \mathbb{E}_q[\mu_1(t)]$, $q_2(1,t) = \mathbb{E}_q[\mu_2(t)]$ and the additional moments:

$$R_{1k}(t) = \mathbb{E}_q[x_k(t)\mu_1(t)]\,, \tag{3.50}$$

$$R_{2k}(t) = \mathbb{E}_q[x_k(t)\mu_2(t)]\,, \tag{3.51}$$

$$R_{12k}(t) = \mathbb{E}_q[x_k(t)\mu_1(t)\mu_2(t)]\,. \tag{3.52}$$

These moments in turn obey the following ODEs (see Appendix B.3):

$$\begin{aligned}
\frac{\mathrm{d}R_{1k}(t)}{\mathrm{d}t} &= -[\lambda_k + g_1(t)]R_{1k}(t) + [A_{k1} + b_k]q_1(1,t) + [A_{k2} + C_k]q_1(1,t)q_2(1,t) + g_{1+}(t)M_{1k}(t) \\
\frac{\mathrm{d}R_{2k}(t)}{\mathrm{d}t} &= -[\lambda_k + g_2(t)]R_{2k}(t) + [A_{k2} + b_k]q_2(1,t) + [A_{k1} + C_k]q_1(1,t)q_2(1,t) + g_{2+}(t)M_{1k}(t) \\
\frac{\mathrm{d}R_{12k}(t)}{\mathrm{d}t} &= -[\lambda_k + g_1(t) + g_2(t)]R_{12k}(t) + [A_{k1} + A_{k2} + C_k + b_k]q_1(1,t)q_2(1,t) \\
&+ g_{1+}(t)R_{2k}(t) + g_{2+}(t)R_{1k}(t)\,,
\end{aligned}$$

where we have defined $g_1(t) = g_{1+}(t) + g_{1-}(t)$ and $g_2(t) = g_{2+}(t) + g_{2-}(t)$. By incorporating all the constraints into the variational free energy functional using Lagrange multipliers, we obtain the following Lagrangian:

$$\begin{aligned}
\mathcal{L} &= KL[q(\nu_1,\nu_2)\|p(\nu_1,\nu_2)] - \langle \log \mathcal{L}\rangle_q \\
&+ \int_0^T \mathrm{d}t\, \xi(t)\left[\frac{\mathrm{d}q_1(1,t)}{\mathrm{d}t} + g_1(t)q_1(1,t) - g_{1+}(t)\right] + \int_0^T \mathrm{d}t\, \zeta(t)\left[\frac{\mathrm{d}q_2(1,t)}{\mathrm{d}t} + g_2(t)q_2(1,t) - g_{2+}(t)\right] \\
&+ \sum_{i=1}^N \int_0^T \mathrm{d}t\, \phi_i(t)\left[\frac{\mathrm{d}M_{1i}(t)}{\mathrm{d}t} + \lambda_i M_{1i}(t) - A_{i1}q_1(1,t) - A_{i2}q_2(1,t) - C_i q_1(1,t)q_2(1,t) - b_i\right] \\
&+ \sum_{i=1}^N \int_0^T \mathrm{d}t\, \kappa_i(t)\left[\frac{\mathrm{d}M_{2i}(t)}{\mathrm{d}t} + 2\lambda_i M_{2i}(t) - 2A_{i1}R_{1i}(t) - 2A_{i2}R_{2i}(t) - 2C_i R_{12i}(t) - 2b_i M_{1i}(t)\right] \\
&+ \sum_{i=1}^N \int_0^T \mathrm{d}t\, \psi_i(t)\Big[\frac{\mathrm{d}R_{1i}(t)}{\mathrm{d}t} + [\lambda_i + g_1(t)]R_{1i}(t) - [A_{i1} + b_i]q_1(1,t) \\
&\qquad - [A_{i2} + C_i]q_1(1,t)q_2(1,t) - g_{1+}(t)M_{1i}(t)\Big] \\
&+ \sum_{i=1}^N \int_0^T \mathrm{d}t\, \gamma_i(t)\Big[\frac{\mathrm{d}R_{2i}(t)}{\mathrm{d}t} + [\lambda_i + g_2(t)]R_{2i}(t) - [A_{i2} + b_i]q_2(1,t) \\
&\qquad - [A_{i1} + C_i]q_1(1,t)q_2(1,t) - g_{2+}(t)M_{1i}(t)\Big]
\end{aligned}$$

$$+ \quad \sum_{i=1}^{N} \int_0^T \mathrm{d}t\, \vartheta_i(t) \Big[\frac{\mathrm{d}R_{12i}(t)}{\mathrm{d}t} + [\lambda_i + g_1(t) + g_2(t)] R_{12i}(t) \\ - [A_{i1} + A_{i2} + C_i + b_i] q_1(1,t) q_2(1,t) - g_{1+}(t) R_{2i}(t) - g_{2+}(t) R_{1i}(t) \Big] ,$$

where we have used Lagrange multipliers $\xi(t)$, $\zeta(t)$ and the set $[\phi_i(t), \kappa_i(t), \psi_i(t), \gamma_i(t), \vartheta_i(t)]$ with $i = [1, 2, \ldots, N]$. Minimisation of the new Lagrange functional with respect to the variational functions ($g_{1\pm}$ and $g_{2\pm}$), the marginal probabilities ($q_1(1,t)$ and $q_2(1,t)$) and the parameters Θ, is obtained with the same procedure described in Section 3.4.1. Functional derivatives and gradients can be found in Appendix B.4.

3.6 Exact inference

As we mentioned above, in a Gaussian-jump process, the joint density $[x(t), \mu(t)]$ is Markovian; therefore we can find an exact solution to the inference problem by using the forward-backward algorithm (Sanguinetti et al., 2009). Here we consider the case with a single Gaussian-jump process, described by Equation 3.1.

Since we are dealing with a nonlinear case, in order to find the moments of the posterior density we need to find the evolution of the marginal posterior $q(x, \mu, t)$. This can be decomposed as (see Section 2.4.5)

$$q(x, \mu, t) = \frac{1}{Z} p(x, \mu, t) \psi(x, \mu, t) , \tag{3.53}$$

where $p(x, \mu, t)$ is the filtering distribution, that is the density of the current state $[x(t), \mu(t)]$ conditioned on the observations up to time t. The function $\psi(x, \mu, t)$ represents the likelihood of future observations, which is the likelihood of the observations after time t given the state $[x(t), \mu(t)]$, and Z is the normalisation constant. Between observations, the filtering density obeys the forward (differential) Chapman-Kolmogorov equation

$$\left[\frac{\partial}{\partial t} + \frac{\partial}{\partial x} (A\mu + b - \lambda x) - \frac{\sigma^2}{2} \frac{\partial^2}{\partial x^2} \right] p(x, \mu, t) = \sum_{\mu' \neq \mu} \left[p(x, \mu', t) f(\mu | \mu') - p(x, \mu, t) f(\mu' | \mu) \right] , \tag{3.54}$$

whereas $\psi(x, \mu, t)$ obeys the backward (differential) Chapman-Kolmogorov equation

$$\left[\frac{\partial}{\partial t} + (A\mu + b - \lambda x) \frac{\partial}{\partial x} + \frac{\sigma^2}{2} \frac{\partial^2}{\partial x^2} \right] \psi(x, \mu, t) = \sum_{\mu' \neq \mu} f(\mu' | \mu) \left[\psi(x, \mu, t) - \psi(x, \mu', t) \right] . \tag{3.55}$$

Observations y_i are included in both the forward and backward messages through the following jump conditions

$$\lim_{s \to t_i^+} p(x, \mu, s) = p(y_i | x(t_i)) \lim_{s \to t_i^-} p(x, \mu, s) , \tag{3.56}$$

$$\lim_{s \to t_i^-} \psi(x, \mu, s) = p(y_i | x(t_i)) \lim_{s \to t_i^+} \psi(x, \mu, s) . \tag{3.57}$$

Since the forward and backward Chapman-Kolmogorov have no analytical solution, they are

solved numerically[9] with the first and last observation ($p(y_1|x(t_1))$ and $p(y_N|x(t_N))$) as initial conditions for the forward and backward equations, respectively.

The product of forward and backward messages gives $Zq(x,\mu,t)$ and the integration of this quantity over the state variables x and μ gives the normalisation constant Z. Therefore, we can compute the marginal posterior $q(x,\mu,t)$ and the following quantities of interest:

$$\begin{aligned} q(\mu,t) &= \int q(x,\mu,t)\mathrm{d}x \\ \langle x(t)\rangle_q &= \int \sum_{\mu} x q(x,\mu,t)\mathrm{d}x\,. \end{aligned}$$

Learning of the parameters $\theta = [A, b, \lambda]$ can be done by minimising $-\log Z$, with respect to the parameters. Z represents the marginal likelihood $Z = p(y_1, \ldots, y_N|\theta)$, which is a function of the parameters.

3.7 Results on a toy dataset

Here we report results of the conditional approximation to the variational approach in the deterministic limit. We first report a comparison of the approximate inference method with the exact inference method developed by (Sanguinetti et al., 2009) and then results on the combinatorial interactions case.

We consider the univariate case with a single telegraph process input. Data are generated by simulating equation

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = A\mu(t) + b - \lambda x(t)\,,$$

with a given input function $\mu(t)$ and a given set of parameters $\theta = [A, b, \lambda]$, and using the following observation model

$$y_i = x(t_i) + v_i \qquad \text{with } \; i = 1, 2, \ldots, N \; \text{ and } \; v \sim \mathcal{N}(0, \sigma^2_{\mathrm{obs}})\,.$$

Figure 3.1 shows inference results obtained by simulating the data with the following input function

$$\mu(t) = \begin{cases} 1 & \text{if } t \in [1, 174] \cup [660, 1000] \\ 0 & \text{if } t \in [175, 659] \end{cases} \tag{3.58}$$

and parameters $A = 3.7 \times 10^{-3}$, $b = 0.8 \times 10^{-3}$, $\lambda = 5.0 \times 10^{-3}$. The variance in the observation model is $\sigma^2_{\mathrm{obs}} = 1.0 \times 10^{-3}$. The left panel shows the posterior marginals $q(1,t)$ obtained with the approximate inference method and the exact inference method, compared to the true input $\mu(t)$. The right panel shows the posterior first moment $M_1(t)$ obtained with the approximate inference method and the exact inference method, versus noisy observations. Parameters θ are not inferred here, but set to to their true values. Figure 3.2 shows the variation in negative

[9] Different numerical approximation schemes can be used to solve the PDE (Vesely, 2001). The results we show are obtained by solving the PDE $\frac{\partial u}{\partial t} + \left[A\frac{\partial}{\partial x} + B\frac{\partial^2}{\partial x^2}\right] u = 0$ by using the following finite differences approximations: $\frac{u_k^{t+1} - u_k^t}{\Delta t} + \left[A\frac{1}{2}\left(\frac{u_{k+1}^t - u_k^t}{\Delta x} + \frac{u_k^t - u_{k-1}^t}{\Delta x}\right) + B\left(\frac{u_{k+1}^t - 2u_k^t + u_{k-1}^t}{(\Delta x)^2}\right)\right] = 0$.

variational free energy during iterations for the approximate inference.

Both exact and approximate method produce a posterior marginal $q(1,t)$ which is comparable with the true input $\mu(t)$, with the exact method performing slightly better than the approximate one.

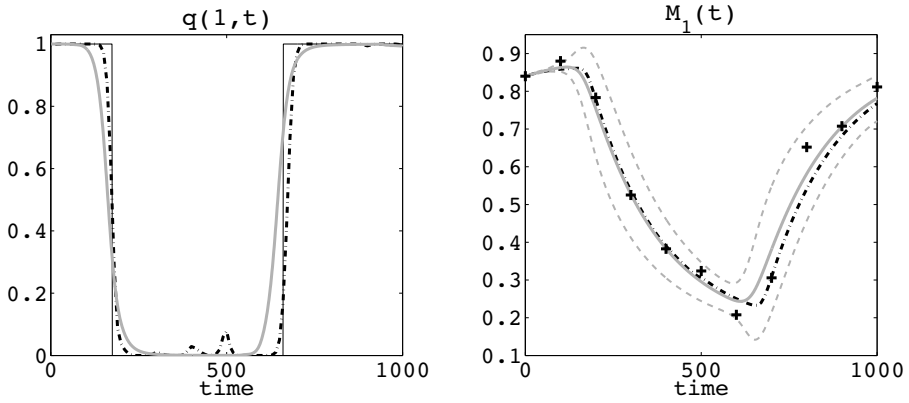

Figure 3.1: Left panel: $q(1,t)$ obtained with exact inference (dotted-dashed) and with approximate inference (solid grey), compared with true input (solid black). Right panel: $M_1(t)$ obtained with exact inference (dotted-dashed) and with approximate inference (solid grey), versus noisy observations (crosses). Confidence intervals for approximate inference (dashed grey), obtained as $M_1(t) \pm 2\sqrt{M_2(t) - M_1^2(t)}$. Confidence intervals for $q(1,t)$ are omitted; the variance of the binary random variable $\mu(t)$ is given by $\sqrt{q(1,t)[1-q(1,t)]}$ (see Bernoulli distribution (Bishop, 2006)).

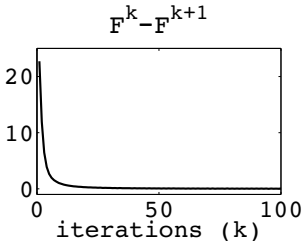

Figure 3.2: Variation of the negative variational free energy: $\Delta[-\mathcal{F}] = [-\mathcal{F}^{k+1}] - [-\mathcal{F}^k]$.

We now consider the multivariate case with two interacting telegraph process inputs. Data are generated by simulating equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{\mu}(t) + \boldsymbol{C}\mu_1(t)\mu_2(t) + \boldsymbol{b} - \boldsymbol{\Lambda}\boldsymbol{x}(t)\,, \tag{3.59}$$

with given input functions $\boldsymbol{\mu}(t) = [\mu_1(t)\ \mu_2(t)]^{\mathrm{T}}$ and a given set of parameters, and using the previous observation model.

Figure 3.3 shows inference results obtained by simulating $N = 3$ processes $x(t)$ with the following input functions

$$\mu_1(t) = \begin{cases} 1 & \text{if } t \in [200, 650] \\ 0 & \text{if } t \in [1, 199] \cup [651, 1000] \end{cases} \qquad \mu_2(t) = \begin{cases} 1 & \text{if } t \in [400, 1000] \\ 0 & \text{if } t \in [1, 399] \end{cases} . \tag{3.60}$$

We used the following set of parameters to simulate the data

$$\boldsymbol{A} = \begin{bmatrix} 3.7 \times 10^{-3} & 0 \\ 0 & 3.7 \times 10^{-3} \\ 3.1 \times 10^{-3} & 0 \end{bmatrix} \qquad \boldsymbol{C} = \begin{bmatrix} 0 \\ 0 \\ 1.3 \times 10^{-3} \end{bmatrix},$$

such that the first process is regulated solely by $\mu_1(t)$, the second process is regulated solely by $\mu_2(t)$ and the third one is regulated by $\mu_1(t)$ and by the interaction of $\mu_1(t)$ and $\mu_2(t)$. Parameters b_i and λ_i are set to $b = 0.8 \times 10^{-3}$ and $\lambda = 5.0 \times 10^{-3}$ for all the processes. The variance in the observation model is $\sigma^2_{\text{obs}} = 1.0 \times 10^{-3}$.

Posterior marginals $q_1(1,t)$ and $q_2(1,t)$ are compared to the true inputs $\mu_1(t)$ and $\mu_2(t)$, respectively (Fig. 3.3, left and centre). Both inferred posterior marginals $q_1(1,t)$ and $q_2(1,t)$ give a good reconstruction of the real inputs $\mu_1(t)$ and $\mu_2(t)$. Figure 3.4 shows the posterior first moments for the three processes, versus noisy observations.

Figure 3.3 (right panel) shows the estimated parameters A_{31}, A_{32} and C_3 for the combinatorially regulated process, versus the true parameter values. Note that a good estimation is provided also for the combinatorial parameter C_3. The estimation improves substantially when the number of observations increases (Fig. 3.3, light grey parts). In fact, in order to have a good inference, but especially a good estimation of the parameters, it is necessary to have a sufficient number of observations for each regulation mode ($\mu_1(t)$, $\mu_2(t)$ and $\mu_1(t)\mu_2(t)$).

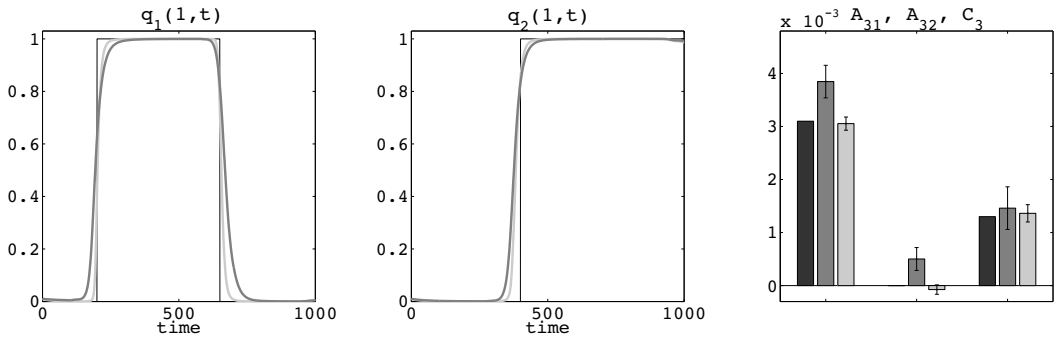

Figure 3.3: Left panel: $q_1(1,t)$ obtained with approximate inference using 10 observations (thick dark grey) and 20 observations (thick light grey), compared with true input (black). Central panel: $q_2(1,t)$ obtained with approximate inference using 10 observations (thick dark grey) and 20 observations (thick light grey), compared with true input (black). Right panel: estimated parameters for $x_3(t)$ using 10 observations (dark grey) and 20 observations (light grey), compared to real values (black).

3.8 Application to *E. coli*'s metabolic data

Here we present an application of the approximate inference method described in Section 3.4 to real gene expression data from a study of *E. coli*'s metabolism. After a short description of the biological context, we show how we used the method to support experimental hypothesis and give useful insights into the metabolism of *E. coli*. Part of the content of this Section has

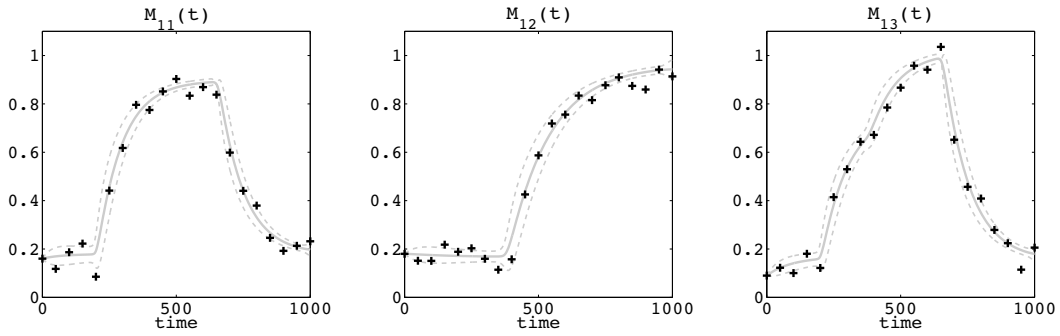

Figure 3.4: Posterior first moments obtained with approximate inference (light grey) versus noisy observations (crosses).

been published in (Rolfe et al., 2012).

3.8.1 Metabolic modes in *E. coli*

The bacterium *E. coli* can live in three possible metabolic modes: aerobic respiration, anaerobic respiration and fermentation. Aerobic respiration is the most efficient one[10], therefore in presence of oxygen is also the preferred one. In absence of oxygen, *E. coli* prefers anaerobic respiration, where it uses nitrate instead of oxygen as final electron acceptor in the electron transport chain. If neither oxygen nor nitrate are present, then fermentation occurs, which is the least efficient metabolic mode.

From the molecular biology point of view, the change from a metabolic mode to another requires a substantial reprogramming of the gene expression, in order for the cell to accommodate different biochemical functions. In particular the remodelling of the gene regulatory dynamics is driven by two oxygen-response agents: the transcription factor FNR (fumarate and nitrate reductase) and the two component system $ArcBA$ (anoxic redox control), whose regulatory part is the transcription factor $ArcA$.

The activation and inactivation of these transcription factors is much faster compared to the time needed for transcription or translation processes[11]. As soon as oxygen is removed from the environment, the transcription factor $ArcA$ becomes active with a simple phosphorylation, while FNR becomes active by forming a dimeric structure. When oxygen is added back to the environment, $ArcA$ is inactivated through a dephosphorylation, while the FNR's dimeric structure becomes unstable and FNR switches to its inactive monomeric form.

Both FNR and $ArcA$ are responsible for the remodelling of the cell metabolism. Once they become active, they switch on the regulatory interactions in the *E. coli*'s transcriptional network such that the aerobic metabolism turns into anaerobic. Analogously, when oxygen becomes present in the *E. coli* environment, the inactivation of FNR and $ArcA$ will switch the transcriptional interactions back to the aerobic mode.

[10]The ratio of ATP molecules produced to the sugar utilised is higher with respect to other metabolic modes.

[11]In *E. coli*, the transition between protein states is of the order $1 - 100\,\mu$s, whereas the time to transcribe a gene and translate a protein are of the order $\sim 1\,\text{min}$ and $\sim 2\,\text{min}$, respectively (Alon, 2006).

3.8.2 Analysis of transcription factor activities in dynamic environments

We are interested in understanding the dynamics of the regulatory interactions occurring during the remodelling from a metabolic mode to another, in particular from aerobic to anaerobic respiration and in the reverse transition. This task requires a detailed knowledge of the dynamics of the activities of multiple transcription factors at the same time, in environments with dynamic oxygen concentration. A measure of these activities in such dynamic environment is hard to be experimentally determined, due to technical constraints. However, a measure of the expression profiles of transcription factors target genes is usually relatively easy through a RT-PCR analysis (Mullis and Faloona, 1987). Therefore statistical models can be used to infer the activities of transcription factors from gene expression data of downstream targets.

Statistical model

Here we use the statistical framework for Gaussian-jump processes to perform this inference analysis. In this context, the process x represents mRNA concentration which is regulated by a transcription factor, whose state is represented by the discrete on/off (representing active/inactive state) variable μ, according to the following system:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}x(t) &= A\mu(t) + b - \lambda x(t) \\ \mu(t) &\sim \mathrm{TP}(f_{\pm})\,. \end{aligned}$$

The binary variable $\mu(t)$ is described by a telegraph process with prior switching rates $f_{\pm}$. The parameters $\theta = [A, b, \lambda]$ can be interpreted as kinetic parameters: A represents the sensitivity of the target gene (or better of its promoter) for the transcription factor, b represents a basal transcriptional rate and finally λ a decay constant which is inversely proportional to the mRNA half life[12]. In case of genes regulated by two transcription factors, the model is given by Equation 3.42 where now $\mu_1(t)$ and $\mu_2(t)$ are the activities of the two transcription factors, A_{ij} is the sensitivity of gene i for transcription factor j and C_i is the sensitivity of gene i for the combinatorial regulation of both transcription factors.

Asymmetricity of transcript profiles

In order to reconstruct the transient dynamics of regulating transcription factors, gene expression of target genes is taken during transitions between aerobic and anaerobic conditions. In the anaerobic-aerobic transition, the initial dissolved oxygen tension raises from 0% to 40% after 1 min and then it becomes stable to 95% after 2 min. In the aerobic-anaerobic transition, the dissolved oxygen tension falls from the initial 65% to 29% after 1 min and stabilises to 0% after 2 min.

In both transitions, high-resolution RT-PCR gene expression data is measured every $\Delta t = 2$ min up to the final time $T = 20$ min, at which the transition is considered concluded and where

[12]The relation between decay constant λ and half life $t_{1/2}$ is: $\lambda = \dfrac{\log 2}{t_{1/2}}$

a new steady state condition is set. At each time step, gene expression is measured as the proportion with respect to the expression at initial time, which for convenience is set to 1 in both transitions.

The rate of change in dissolved oxygen tension is similar and with opposite sign for the two transitions. In other words, during one transition the concentration of dissolved oxygen increases after a few minutes, while in the other transition it decreases after a few minutes. Therefore one may expect that gene expression of genes involved in the remodelling of the metabolism behaves correspondently to this symmetry. In reality, many genes exhibit asymmetrical profiles of abundance in the anaerobic-aerobic and aerobic-anaerobic transitions.

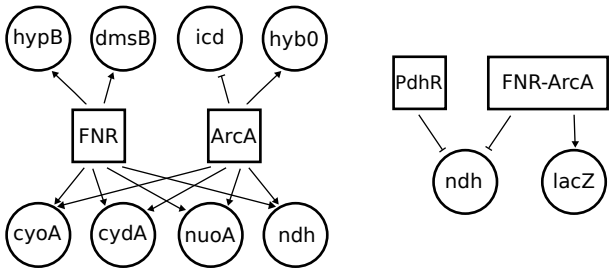

Figure 3.5: Transcriptional network models in *E. coli*. Left: *FNR-ArcA* model. Right: *FNR-ArcA-PhdR* model.

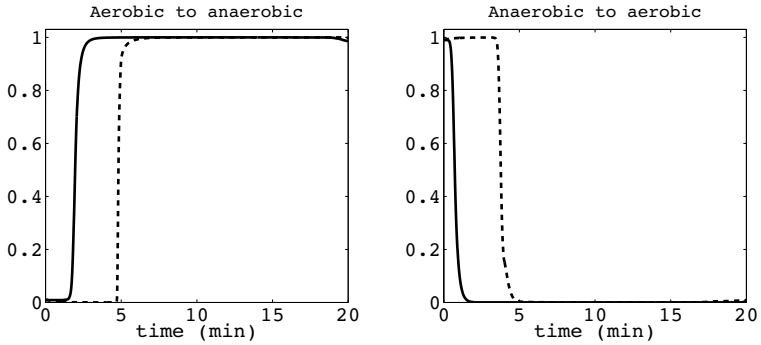

Figure 3.6: Inferred FNR activity obtained with approximate inference using the *FNR-ArcA* model (solid), compared to the one inferred using the reporter gene (dashed).

In this section we look at one of these genes with asymmetrical profile: ndh. It represents a major component in the aerobic respiration and is regulated by both FNR and $ArcA$ transcription factors.

Inference using a *FNR-ArcA* model

In order to find out if the ndh asymmetry is due to the activity of FNR and/or $ArcA$ transcription factors, we used the statistical framework for Gaussian-jump processes to model the ndh profile. By using gene expressions from transcripts regulated by FNR ($hypB$ and $dmsB$), by $ArcA$ (icd and $hyb0$) and by both of them ($cyoA$, $cydA$, $nuoA$ and ndh), we have built the following model

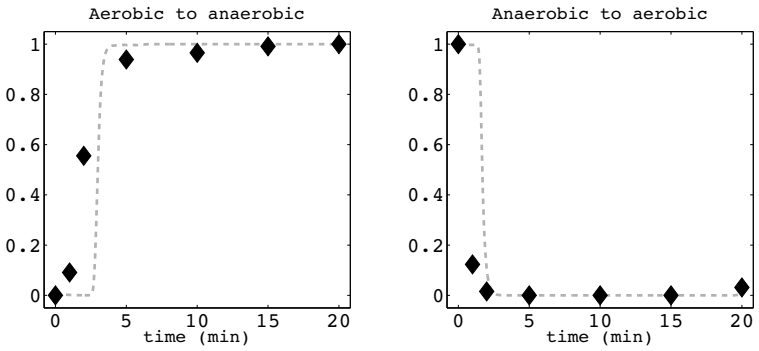

Figure 3.7: Inferred $ArcA$ activity obtained with approximate inference using the *FNR-ArcA* model (dashed), compared phosphorylated $ArcA$ obtained with quantitative densitometry of Western blots (diamonds).

(Fig. 3.5, left):

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{\mu}(t) + \boldsymbol{C}\mu_{FNR}(t)\mu_{ArcA}(t) + \boldsymbol{b} - \boldsymbol{\Lambda}\boldsymbol{x}(t)\,, \tag{3.61}$$

where $\boldsymbol{\mu}(t) = [\mu_{FNR}(t)\ \mu_{ArcA}(t)]^{\mathrm{T}}$ are unknown transcription factor activities. The unknown parameters are defined in the following matrix and vectors:

$$\boldsymbol{A} = \begin{bmatrix} A_{hypB} & 0 \\ A_{dmsB} & 0 \\ 0 & A_{icd} \\ 0 & A_{hyb0} \\ A_{cyoA} & A_{cyoA} \\ A_{cydA} & A_{cydA} \\ A_{nuoA} & A_{nuoA} \\ A_{ndh} & A_{ndh} \end{bmatrix} \qquad \begin{aligned} \boldsymbol{C} &= \begin{bmatrix}0\ 0\ 0\ 0\ C_{cyoA}\ C_{cydA}\ C_{nuoA}\ C_{ndh}\end{bmatrix}^{\mathrm{T}} \\ \boldsymbol{b} &= \begin{bmatrix}b_{hypB}\ b_{dmsB}\ b_{icd}\ b_{hyb0}\ b_{cyoA}\ b_{cydA}\ b_{nuoA}\ b_{ndh}\end{bmatrix}^{\mathrm{T}} \\ \boldsymbol{\Lambda} &= \mathrm{diag}(\lambda_{hypB}, \lambda_{dmsB}, \lambda_{icd}, \lambda_{hyb0}, \lambda_{cyoA}, \lambda_{cydA}, \lambda_{nuoA}, \lambda_{ndh})\,. \end{aligned}$$

The optimisation of the model is performed simultaneously on both the transitions, anaerobic-aerobic and aerobic-anaerobic, by constraining the decay constant of the same genes to be the same in both transitions. Constraints on $\boldsymbol{A}$, $\boldsymbol{b}$ and $\boldsymbol{C}$ are not set, since these parameters are related to the absolute abundance of gene expression level, while we are considering the relative abundance with respect to that at initial time.

Figure 3.6 and Figure 3.7 show the posterior transcription factor activities of FNR and $ArcA$, respectively, inferred using the eight downstream targets mentioned before. The activities of both transcription factors have been both validated in different ways. To validate the inference of FNR, this was compared to the activity of FNR inferred from the expression levels of a reporter $lacZ$ gene (see Appendix B.5) (Fig. 3.6, dashed line). To validate the inference of $ArcA$, a quantitative densitometry of Western blots was obtained to measure the phosphorylation state of $ArcA$ (representing $ArcA$ in active form) during both transitions (Fig. 3.7, diamonds).

Figure 3.8 shows the posterior first moment of all transcripts during both transitions. The

posterior first moments are generally good, especially during the anaerobic-aerobic transition, with the only exception of ndh, whose asymmetric profile cannot be well explained by the model in both the transitions. The profile of $nuoA$ is asymmetric as well, but its gene expression levels are minimal in both the transitions and could be mostly interpreted as noise. On the other hand, ndh exhibits a strong response during the anaerobic-aerobic transition, but not in the aerobic-anaerobic transition.

Inference using a *FNR-ArcA-PhdR* model

The inferred transcription factor activities of FNR and $ArcA$ exhibit symmetric profiles in both transitions (Fig. 3.6 and 3.7). They become inactive or active as soon as the oxygen is added or removed from the environment, respectively. Therefore the asymmetry of ndh profile cannot be attributed to the only control of FNR and $ArcA$, but must be affected by further regulation mechanisms.

In reality, in addition to FNR and $ArcA$, ndh is also regulated by the pyruvate-responsive transcription factor $PdhR$ (pyruvate dehydrogenase complex regulator)[13]. In particular, literature (Ogasawara et al., 2007) and experimental (Rolfe et al., 2012) findings suggests that

- FNR represses ndh 4-fold;
- $ArcA$ activates ndh 2.5-fold;
- $PdhR$ represses ndh 5-fold;
- $ArcA$ and FNR together[14] repress ndh 3-fold.

Like FNR and $ArcA$, $PdhR$ is involved in the *E. coli* oxygen-responsive transcriptional network and like them it responds in both anaerobic-aerobic and aerobic-anaerobic transitions. But, in contrast to FNR and $ArcA$, $PdhR$ is not a direct sensor of the oxygen but an indirect sensor: $PdhR$ is a sensor of pyruvate, which takes some time to accumulate in presence of oxygen. These features make $PdhR$ the perfect candidate as the source of ndh asymmetry. Therefore we included the information about $PdhR$ in the statistical framework, by using the following model

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = A_1\mu_1(t) + A_2\mu_2(t) + C\mu_1(t)\mu_2(t) + b - \lambda x(t)\,, \tag{3.62}$$

where $x(t)$ is the ndh expression and the indices 1 and 2 represent $PdhR$ and the combination FNR-$ArcA$, respectively. We are considering again a combinatorial model where expression of ndh is regulated by two transcription factors: the first is $PdhR$ and the second takes into account of both FNR and $ArcA$ (Fig. 3.5, right). This choice is motivated by the fact that the time shift between the activity profiles of FNR and $ArcA$ during both transitions is negligible, then it makes sense to use a single telegraph process $\mu_2(t)$ to represent both of them.

[13] Pyruvate is a molecule which takes part in the aerobic respiration. During glycolysis, the glucose is degraded into pyruvate, which then enters in the Krebs cycle after its decarboxylation (Berg et al., 2002).

[14] Using an anaerobic $PdhR$ mutant.

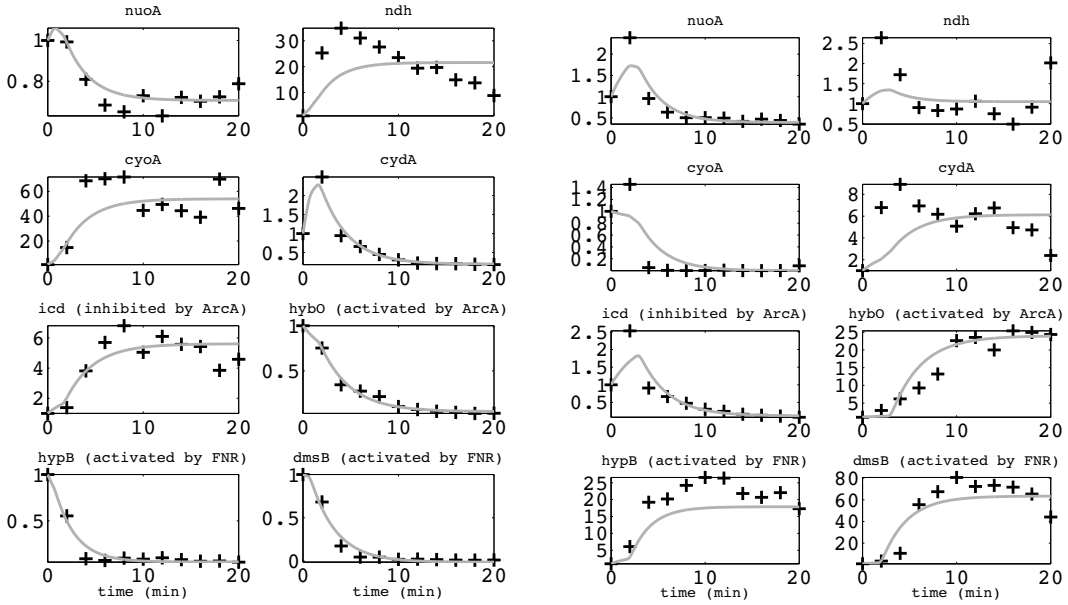

Figure 3.8: Posterior first moments obtained with approximate inference (light grey) versus noisy observations (crosses), during anaerobic-aerobic transition (left) and aerobic-anaerobic transition (right).

During the optimisation, the switching rates of the posterior process approximating $\mu_2(t)$ are set to the ones inferred before for FNR using the reporter gene. By doing this, we optimise only $PdhR$ activity and the parameters A_1, A_2, C, b and λ. Again, we constrain the ndh decay constant to be the same during both transitions and set no constraints for the other parameters.

Posterior first moments, $PdhR$ inferred activity and estimated parameters are showed for both transitions in Figure 3.9. In the anaerobic to aerobic transition all three repressors come off and there is a large initial induction (about 35-fold) of ndh expression. As the cell begins to accumulate pyruvate in presence of oxygen, $PdhR$ starts to rebind and ndh expression decreases back to the starting point. Therefore, in the aerobic to anaerobic transition $PdhR$ is already on the promoter. Further repression by FNR-$ArcA$ has relatively minor effects on the ndh expression, whose level does not change much.

The estimated parameters in the anaerobic-aerobic transition are qualitatively in good agreement with those we reported above from literature: ndh is repressed by both $PdhR$ ($A_1 < 0$) and FNR-$ArcA$ ($A_2 < 0$), with the sensitivity for $PdhR$ higher compared to the one for FNR-$ArcA$ ($|A_1| > |A_2|$). In the reverse transition we do not have similar conclusions, due to the large error bars. Anyway, we can assert that the asymmetric $PdhR$ activity profile during both transitions, can explain the asymmetry in ndh.

To validate the model, predictions of the behaviour of a $PdhR$ mutant were made. The effect of the mutant is that $PdhR$ does not control the transcription of the gene anymore, due to a promoter mutation in ndh; ndh becomes regulated solely by FNR and $ArcA$. Then the prediction is simply obtained by using the estimated parameters and by setting $A_1 = 0$. What we expect in the aerobic-anaerobic transition is that the repression by FNR-$ArcA$ has now a major effect.

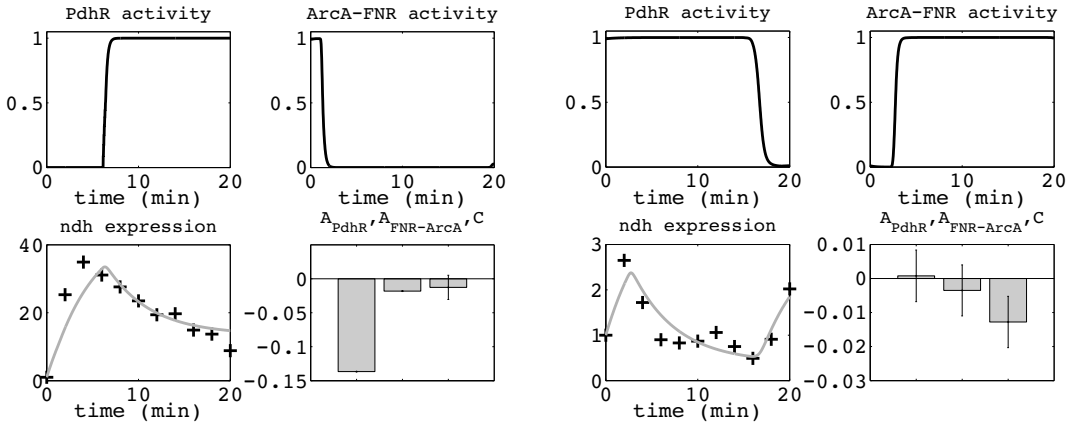

Figure 3.9: Left plots: anaerobic-aerobic transition. Right plots: aerobic-anaerobic transition. Top row: inferred $PdhR$ activities and inferred FNR-$ArcA$ activities. Bottom row: ndh posterior first moment (grey) versus noisy observations (crosses); estimated parameters A_1, A_2, C.

This is showed in Figure 3.10 (left), where the model prediction is compared to the measured $PdhR$ mutant ndh expression. On the other hand, in the anaerobic-aerobic transition, the model cannot explain the speed and the amplitude of the mutant's reaction (Fig. 3.10, right). FNR and $ArcA$ both take a couple of minutes to respond to oxygen addition (as validated before), yet ndh is up 25 fold after two minutes and in the absence of $PdhR$. This suggests that the ndh regulation mechanism may involve something which is not yet fully understood.

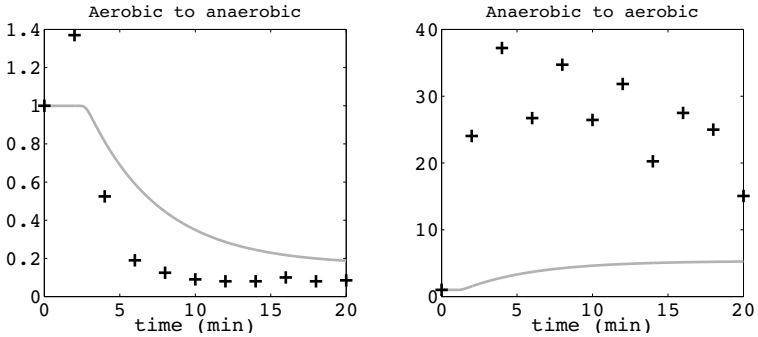

Figure 3.10: Predicted ndh expressions in $PdhR$ mutant (grey) versus noisy observations (crosses).

Chapter 4

Variational inference in feed-forward loops

A knowledge of the dynamics of transcription factors is fundamental to understand the transcriptional regulation mechanism. Nowadays an experimental measure of transcription factor activities in vivo represents a challenge. Several methods have been developed to infer these activities from easily measurable quantities such as mRNA expression of target genes. A limitation of these methods is represented by the fact that they rely on very simple single-layer structures, typically consisting of one or more transcription factors regulating a number of target genes.

Here we present a statistical inference methodology to reverse engineer the dynamics of transcription factors in hierarchical network motifs such as feed-forward loops. The model is based on the deterministic version of the Gaussian-jump process representation given in Chapter 3. Inference and learning problems are solved by using the constrained optimisation algorithm described in Section 3.4, where additional moments are computed by using a Laplace-type approximation and further assumptions.

We demonstrate the method on simulated data and two real datasets. The results on real data show that the predictions of our approach can capture biological behaviours in a more effective way than single-layer models of transcription, and can lead to novel biological insights.

The chapter is divided in six main sections: in Section 4.1, we briefly introduce the problem and our approach; in Section 4.2 we describe in details the model and the inference method. Results on simulated and two real data sets are reported in Section 4.3, 4.4 and 4.5. In Section 4.6, we report conclusions to the chapter.

Most of the material in this chapter is contained in (Ocone and Sanguinetti, 2011).

4.1 Introduction

Unraveling the regulation mechanisms of gene expression is a fundamental problem in systems biology. Recent years have seen tremendous advances towards this goal, driven by technological advances in experimental techniques as well as a systematic application of mathematical modelling. High-throughput techniques, such as microarrays and chromatin immunoprecipitation (ChIP), have uncovered much important information on gene expression profiles and the architecture of biological transcriptional networks. Combining this information in predictive mathematical models can yield precious insights in the way biological systems work.

Nonetheless, in order to understand and predict mechanistically the behaviour of a transcriptional control system, an understanding of the dynamics of transcription factors' (TFs) response to environmental signals is essential. Nowadays, an experimental measure of TFs activities represents a challenge, since TFs are often present at very low concentrations and they are frequently post-transcriptionally regulated through allosteric changes. This motivated many groups to develop quantitative statistical models in order to infer activities of TFs by combining mRNA expression measurements of their target genes with data about the architecture of the regulatory network, usually obtained from ChIP-on-chip experiments.

Broadly speaking, two different classes of approach have been proposed: simplified models of large transcriptional networks, and detailed kinetic models of small subnetworks. The first type of models are usually discrete time models that use log-linear approximations to infer the activities of hundreds of TFs from thousands of target genes (Liao et al., 2003; Sabatti and James, 2006; Sanguinetti et al., 2006). While recent advances (Asif and Sanguinetti, 2011) can model nonlinear interactions between transcription factors on a genome-wide scale, these models are still unable to capture the dynamics of gene expression, and rely on steady-state assumptions. While these models do provide useful insights in biological processes (Partridge et al., 2007; McLean et al., 2010), their simplified nature means that many of the subtleties of regulation will inevitably be lost. The other class of approaches, to which the work in this chapter belongs, adopts a more realistic model of the dynamics of transcriptional regulation based on ordinary differential equations (ODEs) and then infers the profile of the TFs directly from a continuous-time representation of the system. This more faithful representation of the regulatory mechanism however comes at a higher computational cost so that inference in this class of models has so far been possible only on restricted transcriptional networks with simple single-layer architectures. In particular, all methods we are aware of consider simple networks with a single layer of unobserved TFs. In most cases, inference is restricted to the single-input module (SIM) network motif, which is composed of a number of target genes regulated by a single TF (Khanin et al., 2006; Rogers et al., 2007; Lawrence et al., 2007; Gao et al., 2008; Sanguinetti et al., 2009) or at most a few TFs that jointly regulate a number of target genes (Opper and Sanguinetti, 2010). While these simple models constitute a strong proof of principle of the methodology, and may indeed be useful in specific situations (Honkela et al., 2010), many important information processing functions in cells are carried out through hierarchical motifs which entail multiple stages of transcriptional regulation.

In this chapter, we present an ODE-based inference methodology for the most fundamental hierarchical transcriptional network structure, the feed-forward loop (FFL) network motifs. These network motifs consist of a master TF which directly regulates (transcriptionally) a slave TF; both master and slave TF then control the expression of (a number of) target genes, possibly with non-linear interactions at the target promoters. They are frequently encountered in transcriptional regulatory networks due to their robustness and important functions in biological signal processing, such as filtering biological noise fluctuations (Mangan and Alon, 2003). Predicting the dynamics of these fundamental circuits is trivial if we know the parameters of the models and the activity profile of the master TF. Solving the reverse problem of inferring

the master TF activity and model parameters from observations is instead difficult, due to the inevitable non-linearities in these circuits. We use the variational Bayesian approach for approximate inference in continuous-time stochastic processes (Opper and Sanguinetti, 2008), using a telegraph process as prior distribution over the master TF activity (Sanguinetti et al., 2009). The slave TF is assumed to be transcriptionally regulated and we use a logical approximation so that it becomes active when its concentration crosses a critical threshold (inferred from the data).

We test the model extensively on simulated data to assess its identifiability, reporting accurate continuous-time inference of TFs activities, good fitting to data and parameters estimation. We then apply the model on two real data sets: a study of the tumour suppressor protein $p53$ (Barenco et al., 2006), which was previously used as a benchmark for ODE-based inference models (Barenco et al., 2006; Lawrence et al., 2007; Wang and Tian, 2010), and a time-course experiment of *Escherichia coli* undergoing a transition from aerobic to anaerobic environment (Partridge et al., 2007). We show that our approach can be more effective at predicting independent validatory experiments than existing single-layer approaches, as well as being a useful tool for producing novel testable biological hypothesis.

4.2 Model and methods

We consider a FFL consisting of a master TF (whose binary activity state[1] is denoted as $\mu(t)$), a slave TF (whose protein expression we denote as $x_s(t)$) and a target gene whose mRNA expression we denote as $x_t(t)$. The regulation of the target gene is given by a combination of master and slave TFs activities through a logic OR or AND gate. A graphical representation of the network is given in Figure 4.1. Usually, the master TF functions as a sensor for en-

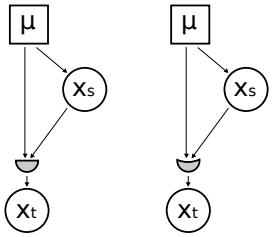

Figure 4.1: FFLs network motifs: OR gate FFL (right), AND gate FFL (left).

vironmental signals; to enable rapid reactions, many TFs have evolved to function as binary switches (active/inactive) triggered by fast post-translational modification (e.g. a phosphorylation) (Alon, 2006). We can capture the switch-like behaviour of the master TF by modelling it as a telegraph process (Sanguinetti et al., 2009). In contrast, we assume that the slave TF is transcriptionally regulated, so that its activity is a (nonlinear) function of its protein abundance. To allow for saturation effects, we model the regulation of the target gene by the slave TF by

[1] A binary representation for both master and slave TF activities is motivated by the fact that transitions between protein states are very rapid (Alon, 2006).

a logical function of the slave TF's protein (Alon, 2006). Mathematically, the model with OR gate is described by the following equations[2]:

$$\frac{\mathrm{d}x_s(t)}{\mathrm{d}t} = A_s\mu(t) + b_s - \lambda_s x_s(t)\,, \tag{4.1}$$

$$\frac{\mathrm{d}x_t(t)}{\mathrm{d}t} = A_t\mu(t) + b_t - \lambda_t x_t(t) + A_c\Theta[x_s(t) - x_c]\,. \tag{4.2}$$

Here Θ represents the Heaviside step function and x_c represents a critical threshold of slave TF protein concentration, such that its regulatory effect becomes important only when its mRNA concentration exceeds x_c. The first equation refers to the regulation of the slave TF protein $x_s(t)$ by the master TF. Here, $\mu(t)$ is the activity of the master TF, given by a telegraph process prior; A_s is the sensitivity of the gene encoding x_s for the master TF; b_s and λ_s represent the basal transcription rate and decay rate of the slave TF, respectively. In practice, we seldom have access to protein measurements of the slave TF, so we will use its mRNA concentration as a proxy for its protein concentration. This is a potentially fraught assumption, as it effectively assumes that the slave TF is transcriptionally regulated and that the protein equilibrates fast compared to the sampling interval in our experimental design. These assumptions are not always met but can be sometimes justified, if only as a rough approximation in the absence of better data. The second equation is similar to the first but it contains an additional term $A_c\Theta[x_s(t) - x_c]$ which takes into account the regulation of the target gene by the slave TF. Transcription of target gene $x_t(t)$ takes into account of the effect of $x_s(t)$ only if $x_s(t) > x_c$; therefore the Heaviside function $\Theta[x_s(t) - x_c]$ represents the binary activation state of slave TF. The model can perform both activation or repression over the three edges of the FFL motif, depending on the value of sensitivity parameters A. In addition it can collapse into a single-input motif (SIM) network when the parameter A_c is null.

In the AND gate FFL, the TF inputs on the target gene are multiplied. In this case, the second equation of the model is described by

$$\frac{\mathrm{d}x_t(t)}{\mathrm{d}t} = A_t\mu(t)\Theta[x_s(t) - x_c] + b_t - \lambda_t x_t(t)\,. \tag{4.3}$$

Also the AND gate FFL can collapse into a SIM if the threshold x_c is null, with $x_t(t)$ becoming a target gene only of the master TF.

Mathematical simulation of the above ODEs is trivial given the profile of the master TF and the model parameters. The problem we wish to address is the reverse problem: given mRNA observations of both the target gene and the slave TF, can we reconstruct both master and slave TF activity profiles and the parameters of the model? This is an analytically intractable problem due to the strong nonlinearity introduced by the Heaviside function in Equation 4.2 and 4.3; we next describe an approximate procedure which can effectively and efficiently handle the problem. Briefly, we compute an approximate variational free energy by introducing two main approximations: the posterior process for the master TF is assumed to be Markovian

[2]Our definition of OR gate is slightly different from the usual Boolean one; we use OR as a synonym of linearly additive response.

(i.e. conditional approximation in Section 3.4), and the induced process on the slave TF is approximated as Gaussian using a Laplace-type approximation[3]. This enables us to derive a gradient for the variational free energy which can then be used in a gradient descent algorithm as showed in Section 3.4.

4.2.1 Inference

The telegraph process representing the activity of the master TF is a continuous-time stochastic process that switches with prior transition rates $f_\pm$ between an ON state and an OFF state. The single time marginal, which represents the probability to be at a given time point in a certain state, is given by the following master equation:

$$\frac{\mathrm{d}p(1,t)}{\mathrm{d}t} = -f_- p(1,t) + f_+ p(0,t)\,, \tag{4.4}$$

$$\frac{\mathrm{d}p(0,t)}{\mathrm{d}t} = -f_+ p(0,t) + f_- p(1,t)\,, \tag{4.5}$$

where $p(1,t)$ and $p(0,t)$ represent the marginal probability that at time t the process is in ON and OFF state respectively.

We assume we have mRNA observations corrupted by i.i.d. zero mean Gaussian noise. In particular the probability to observe the value y_{ti} (with $i = 1,\ldots,N$ and N the total number of observations) if the true value is $x_t(t_i)$ is given by

$$p(y_{ti}|x_t(t_i)) = \mathcal{N}(y_{ti}|x_t(t_i),\sigma_t)\,. \tag{4.6}$$

The same noise model (possibly with different noise variance) is assumed also for the slave TF mRNA observations, y_s. By means of Bayes' theorem we combine the prior distribution over the master TF with the likelihood, in order to compute the posterior distribution of the process $\mu(t)$, given the observations y_s and y_t:

$$p(\nu|y_s,y_t) = \frac{1}{Z} p(y_s|\nu)p(y_t|\nu,y_s)p(\nu|f_\pm)\,. \tag{4.7}$$

This defines the posterior distribution of the master TF activity for all time points as a measure over the space of continuous-time paths of the $\mu(t)$ process[4]. However, as remarked above, direct computation from Equation 4.7 is not possible; we adopt a variational approach to solve the problem. We denote the transition rates of the approximating telegraph process as $g_\pm(t)$, emphasising their dependence on time. It can be shown that the variational approximation is equivalent to minimising the KL divergence of the approximating process $q(\nu)$ with the true posterior $p(\nu|y_s,y_t)$:

$$KL\left[q(\nu)\|p(\nu|y_s,y_t)\right] = \int \mathrm{d}q(\nu)\,\log\frac{q(\nu)}{p(\nu|y_s,y_t)}\,. \tag{4.8}$$

The KL divergence is a functional of the transition rates $g_\pm(t)$ of the approximating process,

[3] See Appendix C.7 for a general description of the Laplace method.

[4] As we defined in Section 3.3, ν represents a continuous-time sample path of the processes $\mu(t)$ over some time interval $[0,T]$.

which in turn determine its single time marginals by the master equation (Eq. 4.4 and 4.5). The state inference problem therefore turns into an optimisation problem (in an infinite dimensional space).

In addition, since the KL divergence is also a function of the model parameters, parameter learning can be easily done through optimisation. Therefore, state inference is carried out in parallel with parameters estimation (see Subsection 3.4.1). It is possible to include prior distributions for the parameters as well to obtain a fully Bayesian (approximate) inference framework; however, in the following we will place a prior distribution only on the critical threshold x_c, for which we define a Gaussian prior distribution centred at half of x_s gene expression.

By plugging Eq. 4.7 into the KL divergence 4.8, we obtain the following expression:

$$KL\left[q(\nu)\|p(\nu|y_s,y_t)\right] = -\left\langle \log\prod_{i=1}^{N} p(y_{si}|x_s(t_i)) + \log\prod_{i=1}^{N} p(y_{ti}|x_t(t_i)) \right\rangle_q + KL\left[q(\nu)\|p(\nu|f_\pm)\right] + \log Z\,. \tag{4.9}$$

The second term on the right hand side of Equation 4.9 represents the KL divergence between two telegraph processes: the KL divergence between two general Markov jump processes has been derived by Opper and Sanguinetti (Opper and Sanguinetti, 2008) and it can be easily computed in terms of process rates and single time marginals of the approximating process. The difficulty lies in computing the likelihood terms in Equation 4.9: these contain the first and second moment of both the variables $x_s(t)$ and $x_t(t)$ under the marginals of the approximating distribution $q(\nu)$.

It is important to remark that an exact minimisation of the KL functional is impossible as it would involve calculating intractable expectations of the Heaviside function $\Theta[x_s(t) - x_c]$ (which arise in the likelihood terms involving the target mRNA x_t). These in principle involve computing all the moments of the slave TF x_s. In practice, we resort to a Laplace-type approximation assuming that $x_s(t)$ is normally distributed (see Subsection 4.2.2).

As we have described in Section 3.4.1, in order to minimise Equation 4.9 with respect to $g_\pm(t)$, it is more convenient to work in an extended space where the KL divergence is considered to be a functional of the single time marginals $q(1,t)$ also. Naturally, process rates and single time marginals are not independent, but they are linked by the master equation (Eq. 4.4 and 4.5). Furthermore, the KL divergence is a functional of the first and second moments of the observed slave gene ($M_{1s}(t)$ and $M_{2s}(t)$) and target gene ($M_{1t}(t)$ and $M_{2t}(t)$) as well. These moments are also related to transition rates $g_\pm(t)$ and single time marginals by a number of ODEs. So the problem turns into a constrained optimisation problem. This can be solved by adding Lagrange multipliers to the KL divergence and performing an efficient gradient descent method based on solving ODEs forward and backward in time.

4.2.2 Heaviside step moments

For the inference in FFL models we have to consider $x_t(t)$ and most of the problems arise from the presence of the Heaviside step function $\Theta[x_s(t) - x_c]$. In order to compute the expectation of $x_t(t)$ under the approximating process, we have to compute the expectation of the Heaviside

step function. This is a nontrivial problem, since the argument of the Heaviside step function contains $x_s(t)$,

$$x_s(t) = e^{-\lambda_s t}\left[x_s(0) + \int_0^t e^{\lambda_s r}\left(A_s\mu(r) + b_s\right)\mathrm{d}r\right], \tag{4.10}$$

which in turn contains the whole history of the stochastic process $\mu(t)$ (Eq. 4.10 is the solution of Eq. 4.1). We address the problem by considering that the Heaviside step function is a deterministic function whose values belong to $\{0, 1\}$ depending on the sign of its argument being negative or positive, respectively. Therefore, the expectation of the Heaviside step function is the cumulative distribution function

$$\langle\Theta[x_s(t) - x_c]\rangle_q = P(x_s(t) > x_c) = \int_{x_c}^{\infty} p(x_s(t))\mathrm{d}x_s(t)\,. \tag{4.11}$$

We can use a Laplace-type approximation for the distribution of $x_s(t)$ giving

$$\langle\Theta[x_s(t) - x_c]\rangle_q \simeq \int_{x_c}^{\infty} \mathcal{N}\left(x_s(t)|M_{1s}(t), M_{2s}(t) - M_{1s}^2(t)\right), \tag{4.12}$$

where $M_{1s}(t) = \mathbb{E}_q[x_s(t)]$ and $M_{2s}(t) = \mathbb{E}_q[x_s^2(t)]$ represent the first and second moment of $x_s(t)$. In other words we have approximated the probability density function of $x_s(t)$, $p(x_s(t))$, as a Gaussian distribution with first moment and variance of the process $x_s(t)$, as its mean and variance respectively. The rationale behind this approximation is that the process $x_s(t)$ can be considered as a sum over many steps of the stochastic process $\mu(r)$ (Eq. 4.10). As a consequence of the central limit theorem, the probability distribution over $x_s(t)$ tends to be Gaussian with the increase of the switching rate.

By means of this approximation, the expectation of the Heaviside step function can be computed through the *error function* and so we can analytically write an ODE for the first moment of $x_t(t)$, $M_{1t}(t) = \mathbb{E}_q[x_t(t)]$, in OR gate FFL:

$$\frac{\mathrm{d}M_{1t}(t)}{\mathrm{d}t} = -\lambda_t M_{1t}(t) + \left(A_t q(1,t) + b_t\right) + \frac{1}{2}A_c\left(1 - \mathrm{erf}(k)\right), \tag{4.13}$$

where $k = (x_c - M_{1s}(t))\left[2\left(M_{2s}(t) - M_{1s}^2(t)\right)\right]^{-\frac{1}{2}}$. For the first moment of $x_t(t)$ in the AND gate FFL, the situation is even more complex since we have to compute the expectation of the quantity $\mu(r)\Theta[x_s(r) - x_c]$. In this case it makes sense to consider the two processes independent, since the Heaviside step function follows the behaviour of $x_s(r)$, whose time scale is described by slower dynamics compared to the switching process $\mu(r)$. This results in the following approximation $\langle\mu(r)\Theta[x_s(r) - x_c]\rangle \simeq \langle\mu(r)\rangle\langle\Theta[x_s(r) - x_c]\rangle$, so we can take advantage of the previous Laplace approximation and obtain the following ODE:

$$\frac{\mathrm{d}M_{1t}(t)}{\mathrm{d}t} = -\lambda_t M_{1t}(t) + A_t q(1,t)\frac{1}{2}\left(1 - \mathrm{erf}(k)\right) + b_t\,. \tag{4.14}$$

Others non trivial expectations are found when we compute the ODE for the second moment of $x_t(t)$. To compute $\langle\mu(r)\Theta[x_s(r') - x_c]\rangle$ (with r and r' two different integration variables), we make the assumption of independence between the process $\mu(r)$ and the Heaviside step func-

tion $\Theta[x_s(r')-x_c]$, when their product is the argument of expectation. Another expectation that is needed is the autocorrelation function of the process $\Theta[x_s(r)-x_c]$, $\langle \Theta[x_s(r)-x_c]\Theta[x_s(r')-x_c] \rangle$. In this case it is possible to show that the value of autocorrelation function for the process $x_s(r)$ decreases exponentially with time interval $r'-r$ and decay constant λ_s (see Appendix C.1). As a consequence, since the Heaviside step function's codomain is defined in the range $\{0,1\}$, we can assume that

$$\langle \Theta_r \Theta_{r'} \rangle = \langle \Theta_r \rangle + \Big(\langle \Theta_r \rangle \langle \Theta_{r'} \rangle - \langle \Theta_r \rangle \Big) \cdot \Big(1 - e^{-\lambda_s (r'-r)} \Big) \,, \tag{4.15}$$

where we have used the short notation Θ_r and $\Theta_{r'}$ for $\Theta[x_s(r)-x_c]$ and $\Theta[x_s(r')-x_c]$, respectively. In simple words it means that autocorrelation is equal to $\langle \Theta_r \rangle$ when $r = r'$, and decreases to the product of the expectations $\langle \Theta_r \rangle \langle \Theta_{r'} \rangle$ as the two Heaviside step functions become uncorrelated. A derivation of the ODEs for the moments and a detailed description of the optimisation algorithm can be found in Appendix C.1 and C.2.

4.3 Results on synthetic data

To benchmark our model and assess the quality of the approximations, we tested the FFL models on simulated data. We use the same experimental set up as in (Sanguinetti et al., 2009). Observations y_s and y_t are given by adding Gaussian noise with standard deviation of 0.03 to 10 discrete time points drawn from the model equations with a given TF activity (input) and known parameters. The master TF is initially on, then switches off and eventually switches on towards the end of the period of interest. The parameters regulating the slave TF, $x_s(t)$, are $A_s = 3.7 \times 10^{-3}$, $b_s = 8 \times 10^{-4}$, $\lambda_s = 5 \times 10^{-3}$. The parameters regulating the target gene are instead the following: $A_t = 2.7 \times 10^{-3}$, $b_t = 8 \times 10^{-4}$, $\lambda_t = 8 \times 10^{-3}$. The additional parameter for the OR gate FFL is $A_c = 2.5 \times 10^{-3}$ and finally we set the critical threshold to a value which allows different regulation areas to contain a sufficient number of observation time points (in this case $x_c = 0.42$). Notice that when we set also the threshold x_c, then the process $\Theta[x_s(t) - x_c]$ is known; therefore it is possible to evaluate the accuracy of the inference not only for the master TF activity $\mu(t)$, but also for the slave TF activity $\Theta[x_s(t) - x_c]$. Parameter estimation is performed in parallel with state inference during gradient descend optimisation. However, the critical threshold x_c is harder to evaluate, as it appears within the non-differentiable function $\Theta[x_s(t) - x_c]$. Therefore we learned the critical threshold off line, by selecting the value that minimises the variational free energy among a set of discrete values used to run different simulations.

Here we report results obtained using the OR gate FFL. Further results on simulated data using the AND gate FFL are reported in Appendix C.2. Figure 4.2 shows the inferred posterior master and slave TF activities, compared with the true TF activities. The posterior slave TF activity is defined as $\langle \Theta[x_s(t) - x_c] \rangle_q$. As can be observed, the method gives an excellent reconstruction of both TF activities, with a precise prediction of the correct transition times and an appropriate switch-like behaviour. Figure 4.3 shows the posterior first moment of x_s

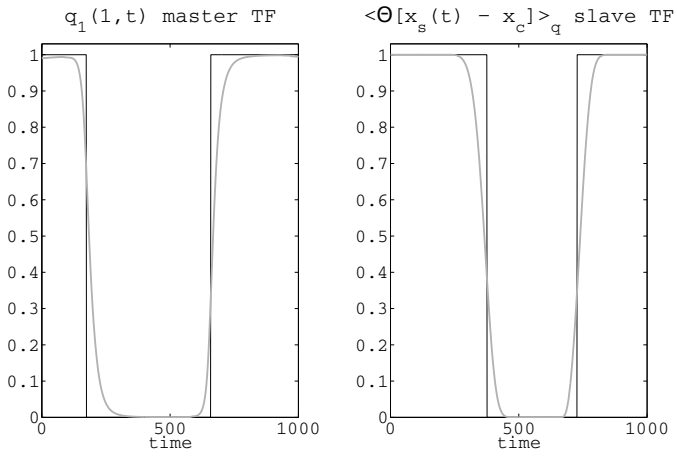

Figure 4.2: Results on simulated data with OR gate FFL. Inferred activity for master and slave transcription factors (thick grey), compared with true inputs (black).

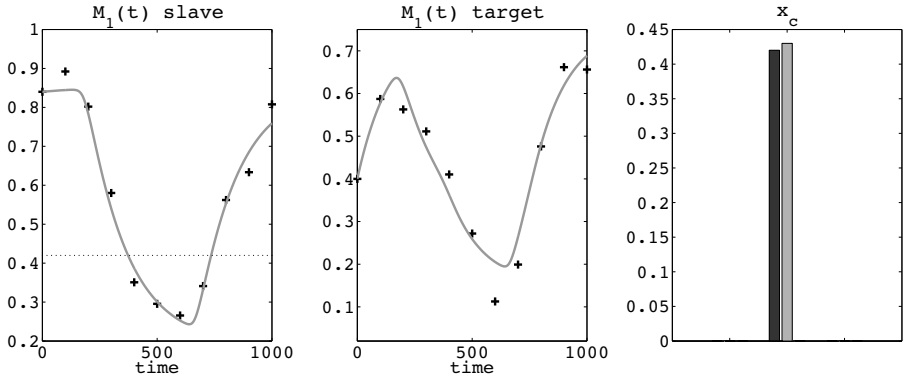

Figure 4.3: Results on simulated data with OR gate FFL. Left and center: posterior first moments (grey) versus noisy observations (crosses) for slave and target gene. Right: estimated x_c (grey), compared to true one (black).

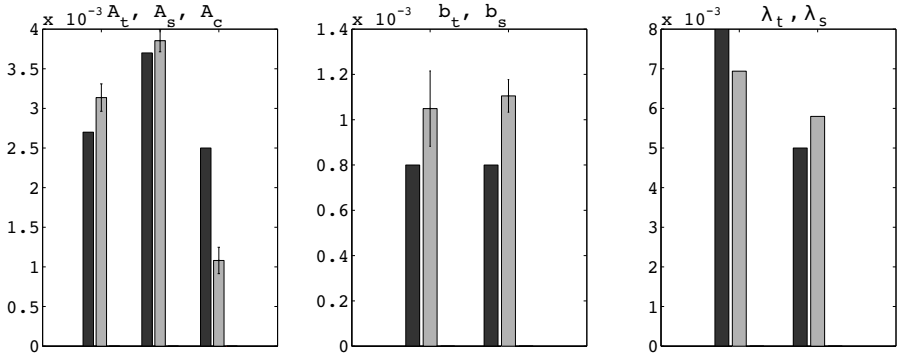

Figure 4.4: Results on simulated data with OR gate FFL. Estimated parameters (grey), compared to true values (black).

and x_t, with observations. Again, the model gives a credible fit to the data, giving confidence that a good optimum of the free energy has been reached. Further results on simulated data and details about the implementation are reported in Appendix C.3. Estimated parameters are reported in Figure 4.4 and 4.3 (right).

4.3.1 Robustness of parameter learning

To assess the robustness of the method, we performed several simulations on different synthetic data sets obtained by changing the values of the kinetic parameters. We created 50 replicates by changing randomly over an order of magnitude the parameter under consideration and letting the rest of the parameters varying randomly over a small range. For example, to assess the estimation of A_s, we created 50 replicates where A_s changes randomly over $[0.0005, 0.005]$, whereas all the other parameters change in a smaller range of values. Figures 4.5 and 4.6 show results of estimation of parameters A_t, A_s and A_c for OR gate FFL and parameters A_t and A_s for AND gate FFL. We plotted the estimated versus the true value of the parameters. The diagonal represents the locus of the points for an ideal estimation. For a quantitative analysis we computed the correlation between estimated and true parameters (Tab. 4.1). In general, the results show a consistent estimation, giving correlation coefficients between the vectors of true and inferred parameters of approximately 0.95 for both OR gate FFLs and AND gate FFLs. Parameter A_c tends to be underestimated in OR gate FFL.

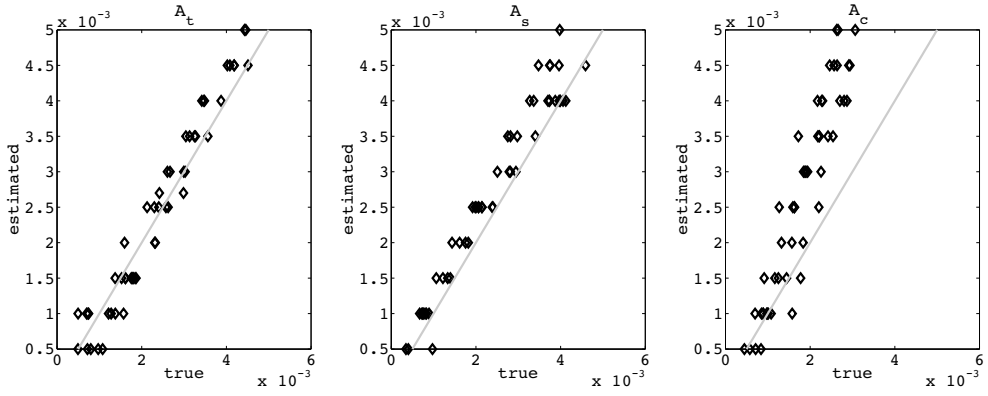

Figure 4.5: Plot of estimated parameter values versus true values in OR gate FFL.

Table 4.1: Correlation between estimated and true parameters.

	OR	**AND**
A_t	0.97	0.98
A_s	0.99	0.93
A_c	0.91	

We also investigated the robustness of the model against mismatch in the noise model. To do so, we generated synthetic data from the model and added Gamma distributed noise, and run

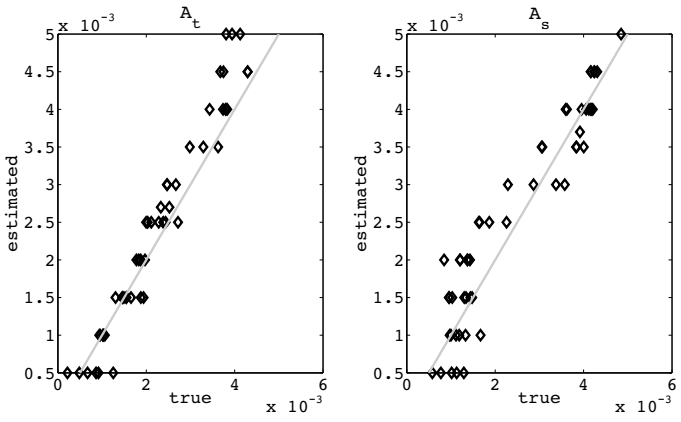

Figure 4.6: Plot of estimated parameter values versus true values in AND gate FFL.

our algorithm (which assumes Gaussian noise). In our experiments, we found that this did not make a significant difference, and the model was still capable of giving a good reconstruction of TF activity (see Appendix C.4).

4.3.2 Comparison with single-input motif model

We compared our FFL model to a SIM model composed of one TF and two target genes. This SIM can be seen as a FFL with a missing interaction between the slave TF and the target gene (Fig. 4.7).

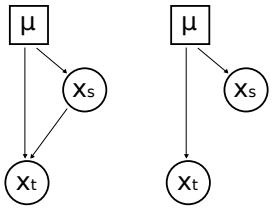

Figure 4.7: FFL and SIM network motifs.

Using the FFL model we generated an artificial data set which is used for inference in both the FFL and SIM models. Results show that the FFL model can explain subtleties which are completely missed by a SIM model (Fig. 4.8). In contrast, as we mentioned above, SIM networks can always be explained by a FFL model: in OR gate FFL, this occurs when $A_c = 0$, whereas in AND gate FFL, when $x_c = 0$.

4.4 Inference of *p53* activity in human leukemia cell line

As a first real data set, we tested our model on the $p53$ prediction task, which has been already tackled by a number of authors with ODE-based models using a SIM architecture (Barenco et al., 2006; Lawrence et al., 2007). This data set focuses on the tumour suppressor protein $p53$. Data were obtained from three independent replicates and represent gene expression at 7

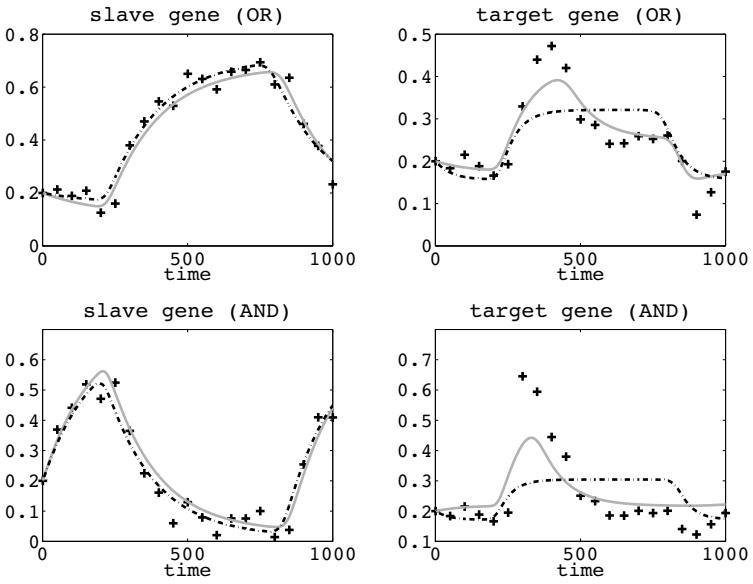

Figure 4.8: Comparison of FFL with SIM on simulated data. Top: OR gate FFL compared to SIM. Bottom: AND gate FFL compared to SIM. Posterior first moment obtained with FFL (solid grey) and with SIM (dashed black), compared to observations (crosses). Parameters for data generated with OR gate FFL: $A_s = 2.5 \times 10^{-3}$, $b_s = 0.8 \times 10^{-3}$, $\lambda_s = 5 \times 10^{-3}$, $A_t = 2.3 \times 10^{-3}$, $b_t = 0.8 \times 10^{-3}$, $\lambda_t = 5 \times 10^{-3}$, $A_c = -2 \times 10^{-3}$, $x_c = 0.53$. Parameters for data generated with AND gate FFL: $A_s = -2.5 \times 10^{-3}$, $b_s = 2.8 \times 10^{-3}$, $\lambda_s = 5 \times 10^{-3}$, $A_t = 6.3 \times 10^{-3}$, $b_t = 0.8 \times 10^{-3}$, $\lambda_t = 5 \times 10^{-3}$, $x_c = 0.30$. Input impulse: $\mu(t) = 0$ if $t \in [0, 499] \cup [1701, 2000]$, $\mu(t) = 1$ if $t \in [500, 1700]$.

discrete time steps from an irradiated human leukemia cell line. After the irradiation, tumour suppressor protein $p53$ increases its activity in order to trigger genes whose function is to protect the cell and eventually induce its apoptosis. An attractive feature of this data set is the existence of a (semi-quantitative) experimental measure of the $p53$ activity, through a Western blot analysis; this allows us to validate, at least qualitatively, the predictions of the model. We compare our predictions with results obtained by Barenco and colleagues using their HVDM model, which presupposes a SIM architecture (Barenco et al., 2006); results using different inference methodologies (but still a SIM architecture) do not differ qualitatively (Lawrence et al., 2007).

Previous models based on a SIM structure could not reproduce the $p53$ experimental measurements (Barenco et al., 2006; Lawrence et al., 2007), predicting a decrease of the $p53$ activity after about 5h time, rather than the experimentally measured 10h time. To explain the discrepancy, an unknown mechanism of migration of a part of the $p53$ protein from the nucleus to the cytoplasm was postulated, but no experimental evidence is available to support this mechanism. While a careful selection of the $p53$ target genes may moderate the problem (Wang and Tian, 2010), we are interested in testing whether the discrepancy may arise from model inadequacy.

In particular, it is known that $p53$ belongs to a modified FFL where another important transcriptional regulator, $E2F1$, controls the expression of $p53$ and directly common $p53$ targets (Polager and Ginsberg, 2009). This represents a FFL where $E2F1$ is the master TF and $p53$

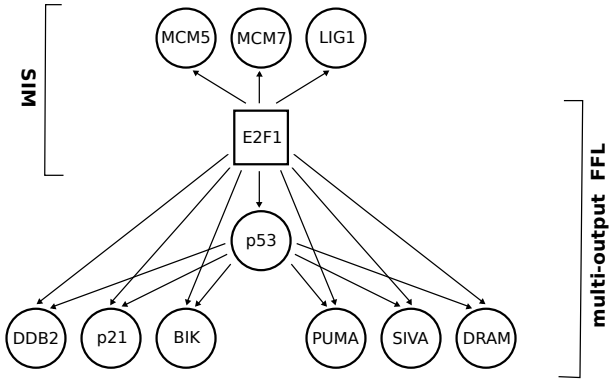

Figure 4.9: $p53$ network architecture. $E2F1$ is the master TF, $p53$ is the slave TF and both regulate target genes *DDB2, p21, BIK, PUMA, SIVA, DRAM*. Target genes of the only $E2F1$ (*MCM5, MCM7, LIG1*) have been included.

is the slave TF (Fig. 4.9).

The target genes are $DDB2$, $p21$, BIK, a subgroup of the genes used in (Barenco et al., 2006) which are regulated by $p53$ and $E2F1$ (Prost et al., 2006; Hiyama et al., 1998; Chinnadurai et al., 2009); in addition we use the apoptotic genes $PUMA$, $SIVA$ and the gene $DRAM$, involved in the process of autophagy, regulated again by both $p53$ and $E2F1$ (Polager and Ginsberg, 2009)[5]. We infer the activity of both $E2F1$ and $p53$ from the observation of the expression of these common target genes and $p53$ itself. We run two different simulations, first using only the SIM part of the model to infer the activity of $p53$ from its target genes and then using the whole FFL model with also another SIM part regarding other genes ($MCM5$, $MCM7$, $LIG1$) regulated only by $E2F1$ (Bracken et al., 2004). In the former case, the $p53$ activity is qualitatively similar to the one predicted by Barenco *et al.* (Fig. 4.10, left). On the contrary, inclusion of the FFL structure leads to a completely different prediction of $p53$ activity which fits much better the experimental measure (Fig. 4.10, right). Details on the experimental platform and further results are reported in Appendix C.5.

Inference of $E2F1$ is also obtained but it is not as interesting as the posterior $p53$ activity, since we do not have an experimental validation for it. Instead, it is interesting that, on this particular real biological data set, both an AND and an OR FFL seem capable of fitting the data reasonably (no strong evidence was found using standard model selection heuristics such as BIC or AIC). Experimental evidence derived from a knock-out of $p53$ shows that the common target genes of $p53$ and $E2F1$ are still activated, indicating that master TF $E2F1$ and slave TF $p53$ cannot be combined into a pure AND gate (Polager and Ginsberg, 2009).

It is important to remark that the assumption that the activity of the slave TF is transcriptionally regulated is violated in this case: in fact, it is known that $p53$ activity is post-translationally regulated by multi-phosphorylation (Lee et al., 2010) and $E2F1$ does not directly affect the transcription of $p53$. Nonetheless, $E2F1$ directly regulates the transcription of other genes that either phosphorylate or interact with $p53$, resulting in a stimulation of its apoptotic activity.

[5]Other apoptotic and non-apoptotic genes mentioned in (Polager and Ginsberg, 2009) have not been included due to their highly noisy expression profiles.

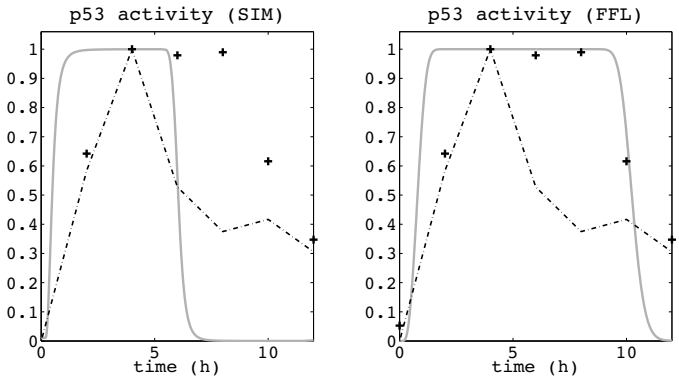

Figure 4.10: Results on $p53$ data set. Left: posterior inferred $p53$ activity using our SIM model (solid grey), compared to Barenco *et al.*'s prediction (dashed black) and experimental measurements (crosses). Right: posterior inferred $p53$ activity using our FFL model (solid grey), compared to Barenco *et al.*'s prediction (dashed black) and experimental measurements (crosses).

Therefore, we may assume that $E2F1$'s activity results in a more efficient activation of $p53$ protein, thus justifying the assumption that slave TF expression is a good proxy for its activity. However, the lack of a direct regulation of $p53$ expression by $E2F1$ means that the physical meaning of the parameters A_s, b_s and λ_s is unclear in this particular case. What our results in fact show is that the inclusion of a hierarchical structure does lead to a considerable improvement in the prediction. Intuitively, this could be explained by the model using the flexibility given by the threshold parameter x_c to adjust the mRNA profile (in this case a suboptimal proxy) to be as close as possible to the true slave TF activity profile.

4.5 Sugar foraging in *E. coli* during aerobic-anaerobic transition

As a further example of the applicability of our methodology, we considered a study of the transcriptional response of *E. coli* during oxygen withdrawal. FFLs are particularly over-represented in bacterial transcriptional networks (Alon, 2006). Furthermore, there is experimental evidence that during transitions *E. coli* adopts a nutrient foraging strategy based on hierarchical transcriptional regulation (Liu et al., 2005). We therefore focus on a subnetwork composed of the cAMP receptor protein (CRP) as master transcriptional regulator, the major regulator of carbon catabolism in *E. coli*. Among the CRP target genes we consider three genes ($manX$, $manY$ and $manZ$) belonging to the PEP-dependent sugar transporting PTS system. The essential function of the $manXYZ$ operon is to make sugar substrates available for metabolism, principally through glycolysis. In addition to the positive regulation by CRP, $manXYZ$ is subject to a negative regulation by mlc, another global regulator of bacterial metabolism. The gene mlc is in turn positively regulated by CRP, leading to the FFL structure shown in Figure 4.12 (Görke and Stülke, 2008). The data set contains *E. coli* global gene expression measurements taken during the transition from aerobic to anaerobic condition, at initial time and four successive discrete time points (5, 10, 15 and 60 minutes) (Partridge et al., 2007).

Using microarray time-courses from target genes $manX$, $manY$ and $manZ$, we infer the

activity of mlc and more interestingly the activity of CRP using an OR gate FFL (details on the experimental platform and further results are reported in Appendix C.6). Posterior expectations give a good fit of the (noisy) target genes expressions (Fig. 4.11, top row); in addition, we note that the inferred parameters A_t and A_s are positive (Fig. 4.11, bottom right), whereas parameter A_c is negative, representing inhibition, which is in accordance with existing knowledge about the sign of regulation in this network.

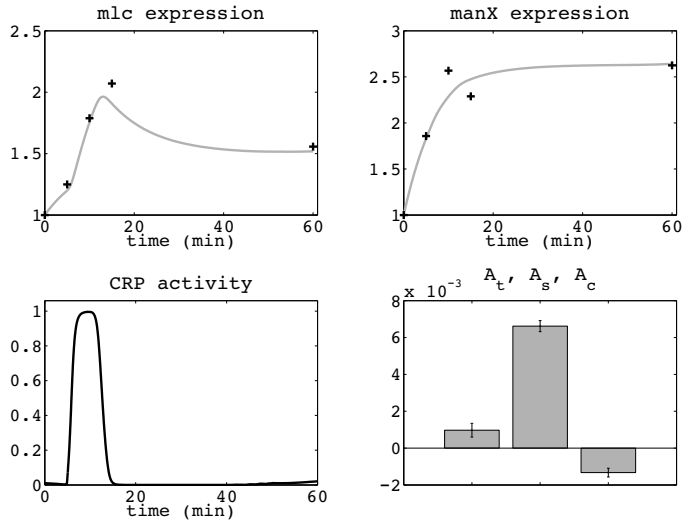

Figure 4.11: Results on *E. coli* data set. Top: posterior inferred mean of *manX* and *mlc* (solid grey) compared to observations (crosses). Bottom: posterior inferred CRP activity (left); estimated parameters of the FFL architecture (right).

In Figure 4.11 (bottom left) we show posterior activity of CRP: it is interesting to note that this activity rapidly transits to the active state and after a short interval it goes back to the inactive state. This behaviour can be rationalised using existing knowledge of the regulation of PEP by the two pyruvate kinases of *E. coli*, whose different temporal behaviour may induce a transient response of CRP. In any case, it directly leads to a testable prediction on the dynamics of cAMP during aerobic-anaerobic transitions, which may lead to insights in the survival strategy of *E. coli* under stress. While this hypothesis clearly needs experimental validation, it gives a useful example of the kind of insights this type of modelling can offer.

4.6 Conclusions

Differential equation models of gene regulation have been enormously successful in the last decade in providing a flexible predictive tool for systems biology. However, their effectiveness depends crucially on the availability of model parameters and external inputs (in the case of non-autonomous systems). Often, these are simply not available, motivating the need for integrating statistical inference tools in ODE-based methodologies. Methodologies for statistical inference in these systems are difficult to develop, but can often offer insights in the underlying biological system which are complementary to those gained by traditional mechanistic

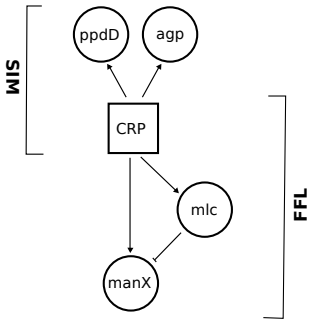

Figure 4.12: CRP network architecture. CRP represents the master TF, mlc represents the slave TF and both regulate target genes *manXYZ* (here only *manX* is showed). Target genes of the only CRP (*ppDD* and agp) have been included too (Keseler et al., 2009).

modelling.

In this chapter, we introduced an approximate inference methodology for ODE-based models of FFLs driven by an unobserved binary TF. While FFLs are very simple structures, they can be viewed as the fundamental building block for hierarchical regulatory networks; from the methodological point of view, there are no significant conceptual difficulties to extending the proposed approach to more complex multi-layer networks. Inference of TF activity in ODE-based models has recently attracted considerable attention, and several tools now exist for network motifs involving a single layer of hidden TFs (Barenco et al., 2006; Lawrence et al., 2007; Rogers et al., 2007; Sanguinetti et al., 2009; Opper and Sanguinetti, 2010). To the best of our knowledge, similar methodologies for hierarchical networks of transcriptional regulators have not been previously reported; given the importance and frequent occurrence of these motifs, the lack of methodology for inference represented a significant obstacle to the adoption of statistical methodologies in ODE based models of biological networks. Our contribution means that inference methods are now available for all of the fundamental building blocks of transcriptional networks. While we view our contribution as mainly methodological, application to two real data sets shows the potential for this type of methodology to yield consistent predictions and to lead to testable hypotheses in biology.

From the methodological point of view, our method belongs to the class of deterministic variational approximate inference techniques which originated in the statistical physics literature and have been widely used in the machine learning community in recent years (Bishop, 2006). While these provide often accurate and efficient estimates of the posterior statistics of interest, they are inherently biased by the choice of family of approximating processes (in our case, Markovian processes). An unbiased alternative would be to resort to stochastic approximations such as Markov chain Monte Carlo. While these have been used successfully in stochastic processes with continuous state space (Wilkinson, 2011), it is not obvious how to derive an efficient sampling strategy for a highly non-linear hybrid model as the one we consider here.

There are several further avenues of research in which this work can be extended. Inference in stochastic models of gene regulation, which are capable of capturing the intrinsic fluctuations

of fluorescence data, are at the moment only available for SIM architectures (Opper et al., 2010), and it would be important to extend them to the FFL case. In Chapter 5 we will present a general framework which can be used to model stochastic FFLs (Ocone and Sanguinetti, 2013). Extrinsic noise sources contributions have also been considered in a Bayesian inference model of gene regulation (Komorowski et al., 2010). Here, a linear noise approximation is adopted to deal with the mesoscopic nature of the observed fluorescent data. For the different levels of gene expression a noise analysis is performed, which would be even more relevant but not straightforward for a hierarchical model such the one presented in this chapter.

Another important direction would be to remove the use of the slave TF mRNA as a proxy for its active protein concentration. This is in principle straightforward in the case when the regulation is purely transcriptional (so that the difference between mRNA and protein profile is entirely due to the difference in half life between protein and mRNA), and can be done by introducing a simple model of translation (Gao et al., 2008).

Chapter 5

Variational inference in Gaussian-jump processes with state-dependent rates

Computational modelling of the dynamics of gene regulatory networks is a central task of systems biology. For networks of small/medium scale, the dominant paradigm is represented by systems of coupled non-linear ordinary differential equations (ODEs). ODEs afford great mechanistic detail and flexibility, but calibrating these models to data is often an extremely difficult statistical problem.

Here we develop a general statistical inference framework for stochastic transcription-translation networks. The model is based on a generalisation of the Gaussian-jump process description given in Chapter 3, which allows an extension to feedback loops.

The inference and learning problems are solved with a variational mean field approximation and results are compared with a generalisation of the exact inference method described in Section 3.6.

We demonstrate the power of the approach on two biological case studies, showing that the method allows a high degree of flexibility and is capable of testable novel biological predictions.

The chapter is divided in eight main sections. Section 5.1 introduces our approach to the problem and Section 5.2 describes the model. In Section 5.3 and Section 5.4 we describe the exact inference method and the variational inference framework. Section 5.5, 5.6 and 5.7 contain results on synthetic data and two real data sets. Conclusion to the chapter are reported in Section 5.8.

Most of the material in this chapter is contained in (Ocone et al., 2013).

5.1 Introduction

Understanding the dynamics of gene regulatory networks (GRNs) is a fundamental area of research in systems biology. *In silico* predictions of the network's response to altered conditions can often give deep insights in the functionality of the biological system under consideration, as well as being crucial in biomedical and biotechnological applications.

Bioinformatics data analysis methods are invaluable in extracting information in large data sets, and can be very useful to predict the main changes in regulatory behaviours (Sanguinetti

et al., 2006; Asif and Sanguinetti, 2011). However, detailed predictions of the dynamics of small/medium scale complex regulatory networks cannot avoid dealing with the non-linear and continuous time nature of such systems, calling for more sophisticated mathematical modelling techniques. By some distance, the dominant paradigm to model GRNs' dynamics is given by systems of coupled non-linear ODEs. ODEs provide an ideal framework for the detailed modelling of mechanistic systems, and of course can rely on refined analysis tools developed over hundreds of years of mathematical research. Nevertheless, mechanistic detail often comes at the cost of including many unknown parameters, as well as novel variables that are not observed (e.g. post-translational modifications of proteins). While there are many parameter estimation tools available (Hoops et al., 2006; Vyshemirsky and Girolami, 2008a; Liepe et al., 2010; Georgoulas et al., 2012), parameter estimation in systems of nonlinear ODEs is often an intrinsically difficult statistical problem due to the severe multimodality of the likelihood landscape. This is further compounded by the limited amount of data usually available in most biological scenarios.

Here we propose a novel statistical modelling framework to model regulatory interactions in GRNs which maintains some key features of nonlinear ODE models while being amenable to a principled statistical treatment. Statistical modelling has become increasingly central in systems biology (Lawrence et al., 2010). Many different statistical models have been proposed in the context of mechanistic systems biology models, ranging from ODEs with uncertain parameters to fully stochastic models (Vyshemirsky and Girolami, 2008a; Wilkinson, 2011). Naturally, the key question is to select a representation which is complex enough to capture the behaviour of the system, but simple enough to allow tractable inference. Here, we build on recently proposed statistical models for transcriptional regulation (Sanguinetti et al., 2009; Opper and Sanguinetti, 2010; Ocone and Sanguinetti, 2011) and represent GRNs using a hybrid continuous/discrete stochastic process, consisting of binary promoter states (occupied/vacant) that drive a stochastic differential equation describing protein dynamics. In this way, we bypass much of the statistical difficulties introduced by detailed modelling of transcription/translation and subsequent post-translational modifications. On the other hand, the introduction of a latent stochastic promoter state can capture much of this complexity, giving a very flexible framework. Our key advance is the introduction of a model of how promoters can depend on upstream protein states, and of a modular approach to approximate inference in this model class that scales linearly with the number of genes in the network. In this way, we can handle medium sized networks of arbitrary topology. We complement our theoretical analysis with an empirical analysis of our method on simulated data, as well as on two real biological systems: the benchmark yeast synthetic network IRMA (Cantone et al., 2009), and the circadian clock of the picoalga *Ostreococcus tauri* (Troein et al., 2011). We compare to existing ODE models, and show that our approach achieves excellent fits and robust predictions. By comparing predictions on different data types, our model also provides a new testable hypothesis about the structure of the *O. tauri* clock network.

5.2 Model

Our aim is to obtain plausible yet statistically tractable models of the dynamics of transcription-translation networks. A central requirement is therefore to include a plausible model of gene expression at the heart of the framework. In this approach, we use the on/off model of gene expression (Ptashne and Gann, 2002), a simple yet powerful model where the rate of transcription of a gene can vary between two levels depending on the occupancy of the promoter of the gene. Assuming for simplicity a tight coupling of transcription and translation, we will use the stronger assumption that protein production can also happen at two distinct rates depending on the occupancy of the promoter. Our network models are therefore composed of a number of connected blocks of two separate types, each of them representing a protein node and a promoter state.

It is convenient to adopt a graphical notation for the statistical models. We denote protein states as circles, and promoter states as squares. Measured protein values are denoted by shaded circles, and we will always assume measurements to occur at discrete times with i.i.d. Gaussian noise; promoter states are assumed not to be observed. Figure 5.1 shows an example of our graphical representation of a 2-gene feedback loop network.

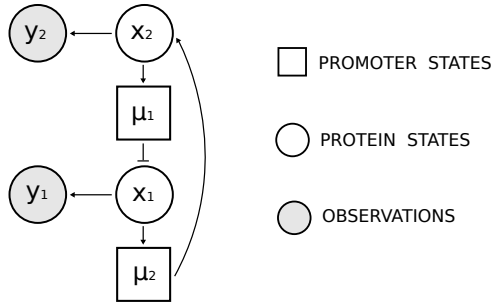

Figure 5.1: Statistical model of a 2-gene feedback loop network.

5.2.1 Promoter model

We model promoters as Markovian continuous time random variables with two possible states, occupied or unoccupied; we denote promoter states as μ and represent them as telegraph processes. The time marginal probability $p_\mu(1,t)$ obeys the chemical master equation

$$\begin{aligned} \frac{\mathrm{d}p_\mu(1,t)}{\mathrm{d}t} &= -f_-(t)p_\mu(1,t) + f_+(t)p_\mu(0,t)\,, \\ \frac{\mathrm{d}p_\mu(0,t)}{\mathrm{d}t} &= -f_+(t)p_\mu(0,t) + f_-(t)p_\mu(1,t)\,, \end{aligned} \tag{5.1}$$

where f_+ and f_- are the switching rates. They represent the transition probabilities per unit time for the switching of the promoter state from 0 to 1 and the other way round, respectively. The time marginal probability $p_\mu(\cdot\,,t)$ represents the probability of the promoter state to have

a certain value (either $\mu = 0$ or $\mu = 1$) at a given time t. For example, the marginal probability $p_\mu(1, t)$ is the probability for the promoter state to be 1 at time t.

Naturally, the rate at which a promoter becomes occupied depends on the state (concentration) of upstream proteins which can bind the promoter. Mathematically, we encode this property by enforcing that the switching rates of the telegraph process μ_i are functions of the TF concentration x_j. As $f_{i\pm}$ represent probabilities per unit time, these functions must be always positive. We use a log-linear model[1] for f_{i+}, primarily due to its mathematical convenience for approximate inference. On the other hand, the switching rate f_{i-} is set to a positive constant value, reflecting the fact that unbinding of the TF (i.e. switch from state 1 to state 0) does not depend on x_j (Schultz et al., 2007). In formulae we have:

$$f_{i+} = k_p \exp\left(k_e x_j\right) , \tag{5.2}$$

$$f_{i-} = k_m , \tag{5.3}$$

where $k_{p,m,e}$ are hyperparameters. Notice that this model implies that the steady state probability of being bound has a saturating, Hill-type dependence on the concentration of protein x_j. By setting to zero the master equation we obtain

$$p_{SS}(\mu = 1|x) = \frac{f_{i+}(t)}{f_{i-}(t) + f_{i+}(t)} = \frac{\exp(k_e x_j)}{\frac{k_m}{k_p} + \exp(k_e x_j)} , \tag{5.4}$$

where $p_{SS}(\mu = 1|x)$ is the steady state probability of the promoter to be in state ON, which depends on the (exponential of the) upstream protein concentration x_j.

5.2.2 Protein model

Protein production is modelled as a stochastic on/off model. We use a continuous approximation to the underlying discrete system and model the transcriptional-translational dynamics through the following stochastic differential equation (SDE):

$$\mathrm{d}x_i = (A_i \mu_i(t) + b_i - \lambda_i x_i)\mathrm{d}t + \sigma \mathrm{d}w(t) , \tag{5.5}$$

where subscript i refers to the target gene and its promoter. Here, $\Theta_i = [A_i, b_i, \lambda_i]$ is the set of kinetic parameters: A_i represents the efficiency of the promoter in recruiting polymerase when occupied. Its sign defines the type of regulation: either activation or repression. Parameter b_i represents a basal transcriptional-translational rate and λ_i is the exponential decay constant for x_i, which is inversely proportional to x_i half-life. Note that Equation 5.5 is a linear SDE conditioned on the history of the promoter state, which entails significant computational efficiency. However, the time-varying nature of the promoter state allows plenty of flexibility to capture non-stationary behaviours. The term $\sigma \mathrm{d}w(t)$, where $w(t)$ is a Wiener process, represents a white

[1] An exponential function is used in order to get a rapid increasing (and decreasing) of the switching rates with the increasing (and decreasing) of the protein concentration. This is essential for the promoter states to have a switching behaviour, which otherwise is not possible (at least empirically) with a linear function.

noise driving process with non-zero variance σ^2. This accounts for the presence of intrinsic noise in the protein concentration x_i, whereas the stochastic process μ_i takes into account of the extrinsic noise in gene expression (Elowitz et al., 2002; Swain et al., 2002).

5.3 Exact inference

As the model of promoter and proteins is jointly Markovian, exact inference[2] can be carried out by numerically solving the Chapman-Kolmogorov forward and backward equations along the lines of (Sanguinetti et al., 2009).

For a 2-gene network, the marginal posterior density can be decomposed as

$$q(x_{1,2}, \mu_{1,2}, t) \propto p(x_{1,2}, \mu_{1,2}, t)\psi(x_{1,2}, \mu_{1,2}, t)\,, \tag{5.6}$$

where $p(x_{1,2}, \mu_{1,2}, t)$ and $\psi(x_{1,2}, \mu_{1,2}, t)$ represent the filtering distribution and the likelihood of future observations, respectively. They obey the forward and backward Chapman-Kolmogorov equation, respectively. Since promoter states are described by two discrete states, we end up with four forward Chapman-Kolmogorov equations

$$\begin{aligned}
\frac{\partial p_{00}}{\partial t} + \sum_{i=1,2}\left[\frac{\partial}{\partial x_i}\left(A_i\mu_i + b_i - \lambda_i X_i\right) - \frac{\sigma^2}{2}\frac{\partial^2}{\partial x_i^2}\right] p_{00} &= f_{1-}p_{10} - f_{1+}p_{00} + f_{2-}p_{01} - f_{2+}p_{00}\,,\\
\frac{\partial p_{01}}{\partial t} + \sum_{i=1,2}\left[\frac{\partial}{\partial x_i}\left(A_i\mu_i + b_i - \lambda_i X_i\right) - \frac{\sigma^2}{2}\frac{\partial^2}{\partial x_i^2}\right] p_{01} &= f_{1-}p_{11} - f_{1+}p_{01} + f_{2+}p_{00} - f_{2-}p_{01}\,,\\
\frac{\partial p_{10}}{\partial t} + \sum_{i=1,2}\left[\frac{\partial}{\partial x_i}\left(A_i\mu_i + b_i - \lambda_i X_i\right) - \frac{\sigma^2}{2}\frac{\partial^2}{\partial x_i^2}\right] p_{10} &= f_{1+}p_{00} - f_{1-}p_{10} + f_{2-}p_{11} - f_{2+}p_{10}\,,\\
\frac{\partial p_{11}}{\partial t} + \sum_{i=1,2}\left[\frac{\partial}{\partial x_i}\left(A_i\mu_i + b_i - \lambda_i X_i\right) - \frac{\sigma^2}{2}\frac{\partial^2}{\partial x_i^2}\right] p_{11} &= f_{1+}p_{01} - f_{1-}p_{11} + f_{2+}p_{10} - f_{2-}p_{11}\,,
\end{aligned}$$

and four backward Chapman-Kolmogorov equations

$$\begin{aligned}
\frac{\partial \psi_{00}}{\partial t} + \sum_{i=1,2}\left[\left(A_i\mu_i + b_i - \lambda_i X_i\right)\frac{\partial}{\partial x_i} + \frac{\sigma^2}{2}\frac{\partial^2}{\partial x_i^2}\right] \psi_{00} &= f_{1+}\left(\psi_{00} - \psi_{10}\right) + f_{2+}\left(\psi_{00} - \psi_{01}\right)\,,\\
\frac{\partial \psi_{01}}{\partial t} + \sum_{i=1,2}\left[\left(A_i\mu_i + b_i - \lambda_i X_i\right)\frac{\partial}{\partial x_i} + \frac{\sigma^2}{2}\frac{\partial^2}{\partial x_i^2}\right] \psi_{01} &= f_{1+}\left(\psi_{01} - \psi_{11}\right) + f_{2-}\left(\psi_{01} - \psi_{00}\right)\,,\\
\frac{\partial \psi_{10}}{\partial t} + \sum_{i=1,2}\left[\left(A_i\mu_i + b_i - \lambda_i X_i\right)\frac{\partial}{\partial x_i} + \frac{\sigma^2}{2}\frac{\partial^2}{\partial x_i^2}\right] \psi_{10} &= f_{1-}\left(\psi_{10} - \psi_{00}\right) + f_{2+}\left(\psi_{10} - \psi_{11}\right)\,,\\
\frac{\partial \psi_{11}}{\partial t} + \sum_{i=1,2}\left[\left(A_i\mu_i + b_i - \lambda_i X_i\right)\frac{\partial}{\partial x_i} + \frac{\sigma^2}{2}\frac{\partial^2}{\partial x_i^2}\right] \psi_{11} &= f_{1-}\left(\psi_{11} - \psi_{01}\right) + f_{2-}\left(\psi_{11} - \psi_{10}\right)\,,
\end{aligned}$$

[2]Note that in this thesis, the term "exact inference" is ambiguous. We use that to refer to an inference method which is not approximate (but exact). In reality, the term "exact" should be used only if one compared with the master equation of all processes treated as discrete.

where, for brevity, we have used the following notation

$$\begin{aligned} p_{s_1,s_2} &= p(x_1, x_2, \mu_1 = s_1, \mu_2 = s_2, t)\,, \\ \psi_{s_1,s_2} &= \psi(x_1, x_2, \mu_1 = s_1, \mu_2 = s_2, t)\,, \end{aligned}$$

and the switching rates f_{1+} and f_{2+} are functions of the protein states x_2 and x_1, respectively[3]. Observation are included in the forward and backward messages as jump conditions

$$\lim_{s\to t_i^+} p(x_{1,2}, \mu_{1,2}, s) = \prod_{j=1,2} \left[p(y_{ji}|x_j(t_i))\right] \lim_{s\to t_i^-} p(x_{1,2}, \mu_{1,2}, s)\,, \tag{5.7}$$

$$\lim_{s\to t_i^-} \psi(x_{1,2}, \mu_{1,2}, s) = \prod_{j=1,2} \left[p(y_{ji}|x_j(t_i))\right] \lim_{s\to t_i^+} \psi(x_{1,2}, \mu_{1,2}, s)\,. \tag{5.8}$$

The marginal posterior probability is then obtained by solving on a grid first the forward coupled PDEs (filtering) and then the backward coupled PDEs (smoothing). The product of the two messages (which has to be normalised) gives the desired marginal posterior. Finally, from the marginal posterior we can compute the quantities of interest

$$\begin{aligned} \langle x_1(t)\rangle_q &= \iint \sum_{\mu_1}\sum_{\mu_2} x_1 q(\mu_1, \mu_2, x_1, x_2, t)\mathrm{d}x_1\mathrm{d}x_2 = \iint x_1 \Big[q_{00} + q_{01} + q_{10} + q_{11}\Big]\mathrm{d}x_1\mathrm{d}x_2 \\ \langle x_2(t)\rangle_q &= \iint \sum_{\mu_1}\sum_{\mu_2} x_2 q(\mu_1, \mu_2, x_1, x_2, t)\mathrm{d}x_1\mathrm{d}x_2 = \iint x_2 \Big[q_{00} + q_{01} + q_{10} + q_{11}\Big]\mathrm{d}x_1\mathrm{d}x_2 \\ q(\mu_1 = 1,\, t) &= \iint \sum_{\mu_2} q(\mu_1 = 1, \mu_2, x_1, x_2, t)\mathrm{d}x_1\mathrm{d}x_2 = \iint \left[q_{10} + q_{11}\right]\mathrm{d}x_1\mathrm{d}x_2 \\ q(\mu_2 = 1,\, t) &= \iint \sum_{\mu_1} q(\mu_1, \mu_2 = 1, x_1, x_2, t)\mathrm{d}x_1\mathrm{d}x_2 = \iint \left[q_{01} + q_{11}\right]\mathrm{d}x_1\mathrm{d}x_2\,, \end{aligned}$$

where q_{s_1,s_2} denotes the normalised product between $p(x_1, x_2, \mu_1 = s_1, \mu_2 = s_2, t)$ and $\psi(x_1, x_2, \mu_1 = s_1, \mu_2 = s_2, t)$.

5.4 Approximate inference

Exact inference requires the numerical solution of a system of high dimensional partial differential equations, leading to severe computational problems when parameters need to be estimated or when more than two genes are present in the network. We therefore adopt an approximate Bayesian approach for the reconstruction of promoter states μ, protein states x and the estimation of model parameters. The quantity we are interested in is the conditional probability distribution $p\,(\mu,\, x|\mathbf{y})$, the joint (posterior) probability of the state of the promoter μ and the promoter concentration x at all time points $t = t_0, \ldots, T$, conditioned on the observations $\mathbf{y}$. We compute an approximation to this quantity by minimising the Kullback-Leibler (KL) functional under a restrictive ansatz for the approximating process. In the follow, we give details for a two-gene network; extension to more genes is straightforward.

[3]We are considering the case of a feedback loop gene network: the promoter state of a gene depends on the protein state of the other gene.

We consider a system summarised as follows:

$$\begin{aligned}
\mu_1(t) &\sim \mathcal{TP}(f_{1\pm}(x_2)) \\
\mathrm{d}x_1 &= (A_1\mu_1 + b_1 - \lambda_1 x_1)\mathrm{d}t + \sigma \mathrm{d}w(t) \\
\mu_2(t) &\sim \mathcal{TP}(f_{2\pm}(x_1)) \\
\mathrm{d}x_2 &= (A_2\mu_2 + b_2 - \lambda_2 x_2)\mathrm{d}t + \sigma \mathrm{d}w(t)\,.
\end{aligned} \tag{5.9}$$

Discrete promoter states are modelled as telegraph processes ($\mathcal{TP}$s), whose switching rates depend on the protein states as described by Equation 5.2 and 5.3. Protein dynamics satisfy SDEs with kinetic model parameters $\Theta_{1,2} = [A_{1,2},\, b_{1,2},\, \lambda_{1,2}]$ and (Wiener) noise driving process $\sigma \mathrm{d}w(t)$.

The choice of the family of approximating distributions $q\,(\mu,\, x)$ is often crucial for computational reasons. Here we build on recent work on approximate inference for continuous-time stochastic processes (Opper et al., 2010) by adopting a structured mean field approximation. By using a mean field approximation, we are assuming that one can ignore the fluctuations in the mRNA. This is not entirely correct since the fluctuations in the mRNA are in reality intimately tied with the fluctuations in the promoter and the assumption is only true for detailed balance condition (Grima et al., 2012). Here, the use of a mean field approximation is mainly justified by mathematical reasons (we want to make tractable the computation of the variational free energy). The approximating distribution then factorises as follows

$$q(\chi_{1,2}, \nu_{1,2}) = q_{\chi_1}(\chi_1)\, q_{\nu_1}(\nu_1)\, q_{\chi_2}(\chi_2)\, q_{\nu_2}(\nu_2)\,, \tag{5.10}$$

where we defined $\chi_{1,2}$ and $\nu_{1,2}$ as continuous-time sample paths of the processes $x_{1,2}(t)$ and $\mu_{1,2}(t)$ over a time interval $[0, T]$. The processes $q_\chi(\chi)$ and $q_\nu(\nu)$ represent pure Gaussian diffusion processes and pure telegraph processes, respectively.

Plugging the approximating distribution into the KL divergence formula we obtain

$$KL[q(\chi_{1,2}, \nu_{1,2}) \| p(\chi_{1,2}, \nu_{1,2} | D_{1,2}, \Theta_{1,2})] = \iiiint \mathrm{d}q_{\chi_1}\, \mathrm{d}q_{\nu_1}\, \mathrm{d}q_{\chi_2}\, \mathrm{d}q_{\nu_2} \log \frac{q(\chi_{1,2}, \nu_{1,2})}{p(\chi_{1,2}, \nu_{1,2} | D_{1,2}, \Theta_{1,2})}\,,$$

where $D_1 = [y_{11}, \ldots, y_{1N}]$ and D_2 represent the discrete protein observations with i.i.d. Gaussian noise. The true posterior distribution is

$$p(\chi_{1,2}, \nu_{1,2} | D_{1,2}, \Theta_{1,2}) = \frac{1}{Z}\, \mathcal{L}\, p(\chi_1|\nu_1)\, p(\nu_1|\chi_2)\, p(\chi_2|\nu_2)\, p(\nu_2|\chi_1)\,, \tag{5.11}$$

where $\mathcal{L}$ is the Gaussian likelihood

$$\mathcal{L} = \prod_{i=1}^{N} p(y_{1\,i}|x_1(t_i)) p(y_{2\,i}|x_2(t_i))\,, \tag{5.12}$$

and the remaining terms in Equation 5.11 are defined by the Equations 5.9. By using the factorised distribution (Eq. 5.10) the KL divergence can be written as the following sum of

terms

$$
\begin{aligned}
KL[q(\chi_{1,2},\nu_{1,2})\|p(\chi_{1,2},\nu_{1,2}|D_{1,2},\Theta_{1,2})] &= \log Z + \sum_{i=1,2}\langle KL\left[q_{\chi_i}(\chi_i)\|p(\chi_i|\nu_i)\right]\rangle_{q_{\nu_i}} \\
&- \sum_{j=1,2}\sum_{i=1}^{N}\langle \log p(y_{j\,i}|x_j(t_i))\rangle_{q_{\chi_j}} \\
&+ \sum_{\substack{i=1,2\\ j=2,1}}\langle KL\left[q_{\nu_i}(\nu_i)\|p(\nu_i|\chi_j)\right]\rangle_{q_{\chi_j}} . \qquad (5.13)
\end{aligned}
$$

The inference problem in this way becomes an optimisation problem, which is solved together with the parameter estimation problem, in a variational expectation-maximisation style. Before describing how we minimise the KL divergence, we report how each of the terms in Equation 5.13 is computed:

1. The first term is constant and can be neglected during the minimisation.

2. The second term contains the KL divergence between two Gaussian diffusion processes, which has been computed in its general form in (Archambeau et al., 2007). Recalling that a Gaussian diffusion process is defined by an SDE with linear drift, the general form for this KL divergence is

$$
KL\left[q\|p_{sde}\right] = \frac{1}{2}\int_0^T \left\langle (\boldsymbol{d}-\boldsymbol{d}_L)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{d}-\boldsymbol{d}_L)\right\rangle_{q_t} \mathrm{d}t\,, \qquad (5.14)
$$

   where q and p_{sde} are an approximating Gaussian process (with linear drift $\mathbf{d}_L$) and a general process (with drift $\mathbf{d}$), respectively, and $\boldsymbol{\Sigma}$ is the covariance matrix for the noise driving process. The expectation inside the integral is computed with respect to the single time marginal distribution q_t.

   In our univariate case, $\mathbf{d} \to A_i\mu_i + b_i - \lambda_i x_i$ and $\mathbf{d}_L \to \alpha_i(t)x_i + \beta_i(t)$ represent the prior drift (conditioned on the upstream promoter state) and the approximating linear function, respectively. The coefficients of the linear approximation, $\alpha_i(t)$ and $\beta_i(t)$, are variational parameters to be optimised. $\boldsymbol{\Sigma} \to \sigma^2$ represents the variance of the noise driving process and $q_t \to q_{\chi_i}$ the single time marginal of our approximating Gaussian process.

   By solving the argument inside the integral in Equation 5.14 and computing the expectation of the KL divergence with respect to q_{ν_i}, we obtain the following expression for the second term:

$$
\begin{aligned}
\langle KL\left[q_{\chi_i}(\chi_i)\|p(\chi_i|\nu_i)\right]\rangle_{q_{\nu_i}} &= \frac{1}{2\sigma^2}\int_0^T \mathrm{d}t\Big[(\alpha_i(t)+\lambda_i)^2[m_i^2(t)+c_i^2(t)] \qquad (5.15)\\
&+ 2(\alpha_i(t)+\lambda_i)\Big[(\beta_i(t)-b_i)m_i(t) - A_i m_i(t) q_{\mu_i}(1,t)\Big] \\
&+ (\beta_i(t)-b_i)^2 + A_i^2 q_{\mu_i}(1,t) - 2A_i(\beta_i(t)-b_i)q_{\mu_i}(1,t)\Big]\,,
\end{aligned}
$$

   where $q_{\mu_i}(1,t) = \langle \mu_i(t)\rangle_{q_{\nu_i}}$ comes from the expectation with respect to the single time

marginal of the telegraph process $q_{\nu_i}(\nu_i)$[4]. The functions $m_i(t) = \langle x_i(t) \rangle_{q_{\chi_i}}$ and $c_i^2(t) = \left\langle x_i^2(t) \right\rangle_{q_{\chi_i}} - \langle x_i(t) \rangle_{q_{\chi_i}}^2$ are the single time marginal mean and variance of the approximating diffusion process $q_{\chi i}(\chi_i)$.

3. The third term contains the likelihood part:

$$\begin{aligned} \left\langle \log p(y_{j\,i}|x_j(t_i)) \right\rangle_{q_{\chi_j}} &\propto \left\langle -\frac{\left(y_{j\,i} - x_j(t_i)\right)^2}{2\sigma_{obs}^2} \right\rangle_{q_{\chi_j}} \\ &\propto -\frac{1}{2\sigma_{obs}^2} \left[y_{j\,i}^2 - 2y_{j\,i} m_j(t_i) + \left(m_j^2(t_i) + c_j^2(t_i) \right) \right] , \end{aligned} \quad (5.16)$$

where σ_{obs}^2 is the observation variance.

4. The fourth and last term can be computed using the variational approximation for telegraph processes (Opper and Sanguinetti, 2008); in addition it requires an expectation with respect to the diffusion process, which involves the prior switching rates (Eq. 5.2 and 5.3). Since only the switching rate f_+ depends on the protein states, we obtain the following equation for the last term:

$$\begin{aligned} \langle KL \left[q_{\nu_i}(\nu_i) \| p(\nu_i|\chi_j) \right] \rangle_{q_{\chi_j}} &= \int_0^T \mathrm{d}t \, q_{\mu_i}(1,t) \left[g_{i-} \log \frac{g_{i-}}{f_{i-}} + f_{i-} - g_{i-} \right] \\ &+ q_{\mu_i}(0,t) \left[g_{i+} \left\langle \log \frac{g_{i+}}{f_{i+}} \right\rangle_{q_{\chi_j}} + \langle f_{i+} \rangle_{q_{\chi_j}} - g_{i+} \right] , \end{aligned} \quad (5.17)$$

where $\langle \cdot \rangle_{q_{\chi_j}}$ denotes expectation with respect to the single time marginal of the diffusion process $q_{\chi j}(\chi_j)$. This expectation can be computed exactly, in fact the diffusion process is approximated with a Gaussian process. Since we have chosen an exponential function for the prior switching rate f_{i+}, the calculation of the expectations in 5.17 under the Gaussian measure q_{χ_j} is trivial (see Appendix D.4.1):

$$\left\langle \log \frac{g_{i+}}{f_{i+}} \right\rangle_{q_{\chi_j}} = \log g_{i+} - \left(\log k_p + k_e m_j \right) , \quad (5.18)$$

$$\langle f_{i+} \rangle_{q_{\chi_j}} = k_p \exp \left(k_e m_j + \frac{c_j^2 k_e^2}{2} \right) . \quad (5.19)$$

The switching rates $g_{i\pm}$ represent the switching rates of the approximating telegraph process, which are other variational parameters to be optimised.

5.4.1 Approximate variational Bayesian scheme

The algorithm for the KL minimisation is based on the iteration of the following three steps:

[4]Note that in Chapter 3.4 (Eq. 3.17) we have already reported a similar result, with some differences. The KL divergence was between two conditional Gaussian diffusion process with prior drift $f(x,\mu) = [A\mu(t) + b - \lambda x(t)]$ and posterior drift $g(x,\mu,t) = [B(t)\mu(t) + d(t) + \alpha(t)x(t)]$, respectively. Here, by using the mean field approximation, we restrict to the case where the posterior diffusion process is a pure Gaussian diffusion process (this means that the variational parameter $B(t)$ is zero at all times). Another consequence of the mean field approximation is the following independence condition: $R = \langle \mu x \rangle_q = \langle \mu \rangle_{q_\nu} \langle x \rangle_{q_\chi}$.

- computation of the approximating diffusion processes;
- computation of the approximating telegraph processes;
- parameter estimation.

Here we describe each of them. For the sake of clarity we remove indices i, j of the genes and consider a single block composed of a promoter state μ and a protein state x, which in turn affects the switching rates of a downstream promoter state μ'.

Approximating Gaussian diffusion process

In the first step we need to compute the approximating diffusion process; under the restrictive assumption of Gaussianity, this is equivalent to computing marginal mean $m(t)$ and variance $c^2(t)$ of the process. Terms involving these statistics are found: in Equation 5.15, which includes the KL divergence between two Gaussian diffusion processes, in the likelihood term 5.16 and finally in Equation 5.17, where the effect of the diffusion is felt on the downstream promoter state. While the first two components correspond to a linear dynamical system which can be minimised by continuous-time Kalman smoothing[5], the third component contributes a term

$$q_{\mu'}(0,t)\left\{-g'_{+}k_e m + k_p \exp\left[k_e m + \frac{c^2 k_e^2}{2}\right]\right\}, \tag{5.20}$$

which is non-linear in the process statistics. Therefore, we cannot use a free form variational inference (i.e. unconstrained) since it would return intractable non-Gaussian diffusions. However, under the assumption of a Gaussian approximation, we can still compute an approximating process by optimising the KL divergence (subject to constraints for the moments) with respect to the variational parameters using an efficient gradient descent algorithm as described below. Note that in Equation 5.20 we have used g'_{+} to denote the posterior switching rates of the downstream promoter μ'.

We recall that the single time marginal mean and variance of an Ornstein-Uhlenbeck process are linked to the drift coefficients by the forward equations

$$\frac{\mathrm{d}m(t)}{\mathrm{d}t} = \alpha(t)m(t) + \beta(t), \tag{5.21}$$

$$\frac{\mathrm{d}c^2(t)}{\mathrm{d}t} = 2\alpha(t)c^2(t) + \sigma^2. \tag{5.22}$$

[5]In this case, minimising the KL divergence means to find an approximating Gaussian process $q_\chi(\chi)$ from a discretely observed conditional Gaussian process $p(\chi|\nu)$ with a Gaussian likelihood for the observations. As we have seen in Section 2.4, this is done by applying the Kalman recursions to the process $p(\chi|\nu)$. The equations for the moments (mean and variance) are propagated forward and then backward, by including the Gaussian observations through jump conditions (Opper et al., 2010). When the process $p(\chi|\nu)$ is not Gaussian, then it is still possible to approximate it with a Gaussian process (Archambeau et al., 2007). But the process depends on all the moments, therefore we cannot use the Kalman recursions anymore. In this case we can minimise the KL divergence (and compute mean and variance of the approximating Gaussian process) using a gradient descent algorithm: this is computationally slower compared to the Kalman recursions, since we need to solve ODEs forward and backward many times, until convergence. Note the possible ambiguity with the use of the term "forward-backward", which we use to refer to the Kalman recursions and to the gradient descent algorithm as well.

Using Lagrange multipliers $\xi(t)$ and $\zeta(t)$, we can incorporate these constraints into the KL divergence functional, obtaining the following Lagrangian (Archambeau et al., 2007)

$$\begin{aligned}\mathcal{L}\left[m, c^2, \alpha, \beta, \xi, \zeta\right] &= \langle KL\left[q_\chi(\chi)\|p(\chi|\nu)\right]\rangle_{q_\nu} - \sum_{k=1}^{N}\langle \log p(y_k|x(t_k))\rangle_{q_\chi} \\ &+ \int_0^T \mathrm{d}t\, q_{\mu'}(0,t)\left\{-g'_+ k_e m + k_p \exp\left[k_e m + \frac{c^2 k_e^2}{2}\right]\right\} \\ &+ \int_0^T \mathrm{d}t\, \xi(t)\left[\frac{\mathrm{d}m}{\mathrm{d}t} - \alpha m - \beta\right] + \int_0^T \mathrm{d}t\, \zeta(t)\left[\frac{\mathrm{d}c^2}{\mathrm{d}t} - 2\alpha c^2 - \sigma^2\right]. \quad (5.23)\end{aligned}$$

This Lagrangian can now be optimised by gradient descent with respect to $\alpha(t)$ and $\beta(t)$. Given a starting estimate of $\alpha(t) = -\lambda$ and $\beta(t) = Aq_\mu(1,t) + b$, we can compute the marginal moments by solving forward in time Equations 5.21 and 5.22 (which is equivalent to setting to zero the functional derivatives of the Lagrangian with respect to the Lagrange multipliers). Taking functional derivatives of the Lagrangian with respect to the moments and setting them to zero, we obtain the following ODEs for the Lagrange multipliers

$$\begin{aligned}\frac{\mathrm{d}\xi}{\mathrm{d}t} &= -\xi\alpha + \frac{1}{2\sigma^2}\left[2m\left(\alpha+\lambda\right)^2 + 2\left(\alpha+\lambda\right)\left(\beta - b\right) - 2A\left(\alpha+\lambda\right)q_\mu(1,t)\right] \\ &+ q_{\mu'}(0,t)\left\{-g'_+ k_e + k_p k_e \exp\left[k_e m + \frac{c^2 k_e^2}{2}\right]\right\} - \frac{1}{\sigma^2_{obs}}\sum_{k=1}^{N}\left[y_k - m(t)\delta(t-t_k)\right], \\ \frac{\mathrm{d}\zeta}{\mathrm{d}t} &= -2\alpha\zeta + \frac{1}{2\sigma^2}\left(\alpha+\lambda\right)^2 + q_{\mu'}(0,t)\left\{\frac{k_p k_e^2}{2}\exp\left[k_e m + \frac{c^2 k_e^2}{2}\right]\right\} + \frac{1}{2\sigma^2_{obs}}\sum_{k=1}^{N}\delta(t-t_k).\end{aligned}$$

These equations are solved backwards from the final condition $\xi(T) = \zeta(T) = 0$. We can then compute the functional gradients of the Lagrangian with respect to the variational parameters

$$\frac{\delta\mathcal{L}}{\delta\alpha} = \frac{1}{2\sigma}\left[2\left(\alpha+\lambda\right)\left(c^2+m^2\right) + 2m\left(\beta - b\right) - 2Amq_\mu(1,t)\right] - m\xi - 2c^2\zeta, \quad (5.24)$$

$$\frac{\delta\mathcal{L}}{\delta\beta} = \frac{1}{2\sigma}\left[2\left(\beta - b\right) + 2m\left(\alpha+\lambda\right) - 2Aq_\mu(1,t)\right] - \xi, \quad (5.25)$$

and perform a gradient step in a gradient descent.

In practice, the contribution from Equation 5.20 to the Lagrangian is often negligible compared with the contribution from the observations. The omission of the term 5.20 would allow approximate minimisation using continuous time Kalman smoothing at a fraction of the computational cost. To obtain the results we show in Section 5.5, we used a Kalman smoother as an initialisation for the gradient descent procedure, which usually converged after very few steps.

Approximating two-state Markov jump process

In the second step, we compute the approximating jump process marginals $q_\mu(1,t)$ and rates $g_\pm(t)$. Inspection of the KL divergence reveals that the marginals are only involved linearly (Eq. 5.15 and Eq. 5.17), so that fast forward-backward recursions can be used for these computations (Opper et al., 2010).

Since the process marginals are linked to the switching rates through the master equation, the minimisation of the KL divergence represents a constrained optimisation problem. The

master equation is then included as a constraint through a Lagrange multiplier $\psi(t)$. By assuming that the switching rates of the promoter state μ are affected by an upstream protein state x', the resulting Lagrangian is

$$\begin{aligned}\mathcal{L}[q_\mu, g_\pm, \psi] &= \langle KL\left[q_\nu(\nu)\|p(\nu|\chi')\right]\rangle_{q_{\chi'}} + \int \mathrm{d}t \frac{1}{2\sigma^2}[A^2 - 2A(\alpha+\lambda)m - 2A(\beta-b)]q_\mu(1,t) \\ &+ \int \mathrm{d}t\, \psi(t)\left(\frac{\mathrm{d}q_\mu(1,t)}{\mathrm{d}t} + (g_- + g_+)q_\mu(1,t) - g_+\right), \end{aligned} \tag{5.26}$$

where we have included only the terms of interests, which depend on the process $q_\mu(1,t)$. Compared to the optimisation in (Opper et al., 2010), in our trancriptional-translational model we do not have a simple KL divergence term $KL\left[q_\nu(\nu)\|p(\nu)\right]$ but its expectation with respect to a diffusion Gaussian process. By setting to zero the functional derivatives of $\mathcal{L}$ with respect to the posterior switching rates $g_\pm(t)$, we obtain the update formula for the posterior switching rates: $\log g_\pm = \langle \log f_\pm \rangle_{q_{\chi'}} \pm \psi$ (see Appendix D.4.2). By using the relations 5.3 and 5.18, these updates become

$$\begin{aligned} g_+ &= k_p \exp(k_e m') \exp(+\psi), \\ g_- &= k_m \exp(-\psi), \end{aligned}$$

where we have used $m' = \langle x' \rangle_{q_{\chi'}}$ to denote the single time marginal mean of the approximating diffusion process for the upstream protein x'. Then, by setting to zero the functional derivative of the Lagrangian $\mathcal{L}$ with respect to $q_\mu(1,t)$, we obtain an ODE for the Lagrange multiplier ψ:

$$\begin{aligned}\frac{\mathrm{d}\psi}{\mathrm{d}t} &= \left[g_- \left\langle \log \frac{g_-}{f_-} \right\rangle_{q_{\chi'}} + \langle f_- \rangle_{q_{\chi'}} - g_-\right] - \left[g_+ \left\langle \log \frac{g_+}{f_+} \right\rangle_{q_{\chi'}} + \langle f_+ \rangle_{q_{\chi'}} - g_+\right] \\ &+ (g_- + g_+)\psi + \frac{1}{2\sigma^2}[A^2 - 2A(\alpha+\lambda)m - 2A(\beta-b)]. \end{aligned} \tag{5.27}$$

Introducing a new variable $r = \exp(-\psi)$ and substituting the new switching rates

$$g_+ = k_p \exp(k_e m')\, r^{-1}, \tag{5.28}$$

$$g_- = k_m\, r, \tag{5.29}$$

into the ODE for ψ, we obtain the following ODE for r (see Appendix D.4.3):

$$\frac{\mathrm{d}r}{\mathrm{d}t} = k_p \exp(k_e m')\left[r\, \exp\left(\frac{1}{2}c'^2 k_e^2\right) - 1\right] - k_m\, r(1-r) - \frac{1}{2\sigma^2}[A^2 - 2A(\alpha+\lambda)m - 2A(\beta-b)]r, \tag{5.30}$$

where $c'^2 = \left\langle x'^2(t) \right\rangle_{q_{\chi'}} - \langle x'(t) \rangle^2_{q_{\chi'}}$. Solving backward this equation (with final condition $r(T) = 1$, i.e. final condition for the Lagrange multiplier $\psi(T) = 0$), provides the value of r at all times. Using $r(t)$ we can update the switching rates $g_\pm(t)$ (Eq. 5.28 and 5.29) and then the single time marginals $q_\mu(1,t)$ (through the master equation).

Learning of kinetic parameters

In order to update the kinetic parameters $\Theta = [A,\, b,\, \lambda]$, we have to minimise the KL functional with respect to them. In the KL divergence (Eq. 5.13), the term which depends on Θ is the one given by Equation 5.15. Also the likelihood term (Eq. 5.16) depends on the kinetic parameters, but only implicitly; therefore it is disregarded during the minimisation[6].

By inspection of Equation 5.15, we see that it is quadratic in A, b and λ, so its minimisation represents a simple quadratic programming problem. By defining $f = \langle KL\left[q_\chi(\chi) \| p(\chi|\nu)\right]\rangle_{q_\nu}$ and $x = [A\ b\ \lambda]^{\mathrm{T}}$, we have that $f = \frac{1}{2}x^{\mathrm{T}}Hx + G^{\mathrm{T}}x$, where H and G represent the Hessian and the gradient of f, respectively. By setting to zero the derivative of the cost function f with respect to x, we obtain[7] the updated vector for the kinetic parameters:

$$\frac{\mathrm{d}f}{\mathrm{d}x} = Hx + G = 0 \quad \longrightarrow \quad x = -H^{-1}G\,. \tag{5.31}$$

This provides an estimation for the mean of each parameter. In addition, assuming the parameters are Gaussian distributed, we also get an estimation of the variance of each parameter, which is simply given by the diagonal elements of the inverse Hessian matrix H^{-1} (Yuen, 2010). This information provides a confidence interval for the parameter estimation and can be used to evaluate statistically the goodness of the estimation.

Hyperparameters in the switching rates, as well as the system noise, are fixed to reasonably vague values (their precise identifiability would require longer time series than usually available).

5.5 Results on synthetic data

To benchmark our approach, we tested our hybrid regulatory model on data simulated from a two-genes negative feedback loop which exhibits oscillatory dynamics. The data were generated in the following way: given a set of parameters $\Theta_{1,2} = [A_{1,2},\, b_{1,2},\, \lambda_{1,2}]$ for the two genes, we simulated the model to produce continuous time courses for both promoters and proteins. Noisy observations at discrete time points were obtained from the protein time courses by adding i.i.d. Gaussian noise. We simulated the model across many runs to empirically assess the quality of our results; in particular, we focused on two issues, quality of the variational approximation and accuracy of the estimation of the parameters.

To probe the first issue, we compared profiles reconstructed using our variational approach with exact inference results. Given the high computational costs of exact inference (>2 minutes for a single run) and the focus on the quality of the posterior approximation, we compared the two approaches with parameters fixed to the true values. Figure 5.2 (left and centre) shows the two reconstructed promoter states in a specific example; as can be seen, the approximate

[6] The likelihood term depends on the moments of the diffusion process (m and c^2) which in turn depend on the kinetic parameters. After the update of the diffusion and the jump processes, we are in the situation where the KL divergence has reached its minimum with respect to m and c^2, therefore the functional derivatives of the KL with respect to m and c^2 are zero.

[7] We recall the following rules of derivation: $\frac{\mathrm{d}}{\mathrm{d}x}\left(x^{\mathrm{T}}Hx\right) = 2Hx$ if $H = H^{\mathrm{T}}$; $\frac{\mathrm{d}}{\mathrm{d}x}\left(G^{\mathrm{T}}x\right) = G$.

results are very close to those inferred with the exact method.

Figure 5.2 (right) shows a quantile plot obtained by estimating parameters from ten data sets with random parameter values. The parameters range is confined to a certain region of the parameter space in order to allow the system to exhibit oscillations; however, the simulated oscillatory data still spans a range of different amplitudes and frequencies. Given that our method provides an estimate of the posterior variance of each parameter, we can quantify exactly how many true values fall below each predicted quantile. For each quantile on the x-axis, the value on the y-axis shows the number of estimated parameters (normalised to the total number of estimated parameters) which belong to that quantile. An ideal estimation should produce a diagonal $x = y$ line, showed in grey in Figure 5.2 (right). In our case, the figure reveals a reasonable accuracy, with a slight underestimation of the posterior variance, which is a widely acknowledged problem of mean field variational methods (Opper and Winther, 2001).

In terms of computational time, the whole inference procedure is very efficient, giving the above results in a few seconds. Furthermore, the mean field approach has complexity that scales linearly with the number of genes/promoters, guaranteeing at least in principle the applicability of the method to larger systems.

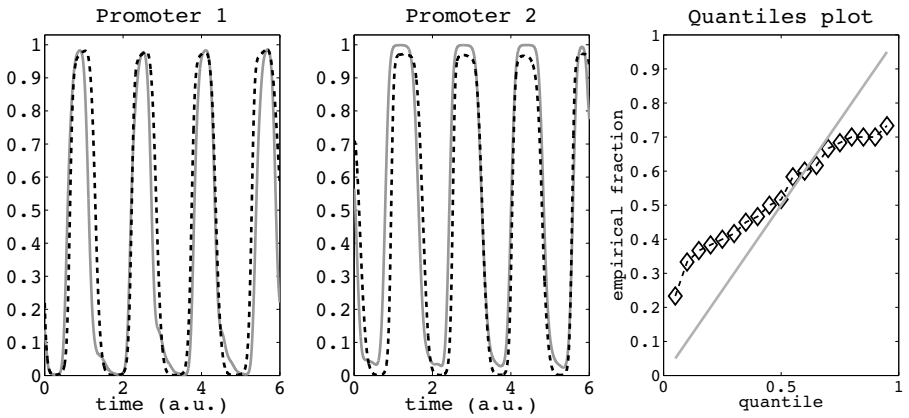

Figure 5.2: Results on simulated data. Left and centre: posterior inference of promoter states using the variational method (dashed black), compared to reconstruction with exact inference (solid grey). This example was obtained with the following set of parameters: $A_{1,2} = [0.0036, -0.0031]$, $b_{1,2} = [0.00025, 0.0033]$, $\lambda_{1,2} = [0.0025, 0.0025]$. Right: quintile plot representing the goodness of parameter estimation across many runs.

In the next two sections we assess the performance of the hybrid regulatory model on two real data sets. The main features we are interested in are the quality of the fits to the training data (i.e. whether the model has sufficient flexibility to capture the complex behaviour of biological circuits) and the ability to predict unseen data in perturbed conditions (i.e. whether it is able to generalise).

5.6 Modelling the IRMA synthetic yeast network

As a first application, we considered the IRMA network (Cantone et al., 2009), a synthetic network embedded in the yeast *Saccharomyces cerevisiae*. IRMA is composed of five genes: $ASH1$, $CBF1$, $GAL4$, $SWI5$ and $GAL80$. Figure 5.3 shows a representation of our hybrid regulatory model for the IRMA network, where the interactions between the five genes can be easily detected by looking at the thick black lines. The network was engineered to respond to changes in the sugar supplied (galactose vs glucose). Gene expression from all the five genes was measured during the transitions from glucose to galactose and from galactose to glucose, giving two sets of data which are referred to as switch-on and switch-off time series.

To analyse the dynamics of the IRMA network we compared two different models: our hybrid regulatory model and the nonlinear delay differential equation (DDE) model of Cantone *et al.* (Cantone et al., 2009). Our model consists of five SDEs with 3×5 free parameters, while the model of Cantone and colleagues consists of five DDEs, modelling the mRNA levels of the five genes, with a total of 31 parameters. Both models were trained by using only the switch-on time series: for the hybrid regulatory model we adopt the variational Bayesian scheme described above, whereas the Cantone *et al.*'s model is trained using stochastic optimisation (Appendix D.3). Figure 5.4 shows the results of this analysis. The left hand column shows the fit to the training switch-on data: the black lines represent the fits of the Cantone *et al.*'s model, while the grey lines are the hybrid regulatory model posterior predictions (with confidence intervals). Both models give a qualitatively good fit, with a slightly better fit for the hybrid regulatory model. The right hand column in Figure 5.4 shows the simulated switch-off behaviour obtained using the parameters estimated from the switch-on transition data. Both models capture the general de-activation trend, but the hybrid regulatory model seems to give a slightly better prediction of the initial transient behaviour of $ASH1$.

5.7 Modelling circadian clock in *O. tauri*

Despite its complex topology and the relatively large number of genes involved, the IRMA network does not exhibit particularly complex dynamics during the two transitions. As a second example therefore we consider a circadian clock, i.e. a network which can sustain oscillatory dynamics autonomously. By standard results in dynamical systems theory, this implies the presence of feedback loops in the network architecture[8]. Typically, transcription-translation models are used to explain the sustained oscillations of gene expression. In a minimal model, the translational product of a clock gene becomes the transcriptional activator/inhibitor of another clock gene. Given the importance of circadian clocks for biomedical applications, and the availability of many tools to study oscillatory time series in mathematics and engineering, circadian clocks have become a major focus of systems biology research.

The picoalga *Ostreococcus tauri* has recently emerged as a powerful yet simple model of plant circadian clocks due to its compact genome and extremely simple physiology. Notably,

[8]This is true with a purely deterministic model. In a stochastic model, as the one presented in this chapter, this is not necessarily true. Noise can induce oscillations in models with no feedback loops (Toner and Grima, 2013).

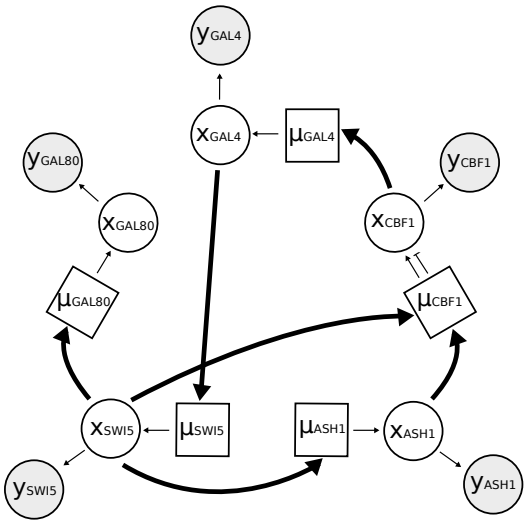

Figure 5.3: IRMA yeast synthetic network. Thick black lines, which model promoter activations, show the IRMA network topology. Activation and repression arrows in the transcriptional-translational dynamics between μ_{CBF1} and x_{CBF1}, model the fact that $ASH1$ and $SWI5$ are a repressor and an activator of $CBF1$, respectively.

only the clock genes $TOC1$ and $CCA1$, represented by multiple members in the higher plant *Arabidopsis thaliana*, are encoded in *O. tauri*, along with a cryptochrome-like gene with possible clock involvement (Heijde et al., 2009). This has led to the hypothesis that its clock network consists of a minimal oscillator of a single loop between these two genes. This hypothesis has been explored mathematically in a number of papers (Morant et al., 2010; Thommen et al., 2010; Troein et al., 2011); most recently, Troein *et al.* provided a comprehensive data sets consisting of several luciferase time series measurements of $TOC1$ and $CCA1$ protein abundance in a synchronised population of *O. tauri* cells. Troein and collaborators then proposed a detailed ODE model of the system, and used 144 luciferase time series to parameterise the model.

We compare the results of our hybrid approach with the ODE approach of Troein *et al.* on the *O. tauri* circadian clock data. The structure of the model is a simple negative feedback loop (NFL) network, including the evening gene $TOC1$ and the morning gene $CCA1$ (Fig. 5.9A). We consider the presence of a single light input affecting the $CCA1$ promoter state. This is to mimic light-induced phosphorylation of the $TOC1$ transcription factor (Troein et al., 2011), which affects its ability to bind the $CCA1$ promoter (Appendix D.1).

In order to evaluate the performance of our approach, we compare our stochastic hybrid approach and the complex clock model of (Troein et al., 2011). Our hybrid regulatory model only has 2 SDEs and 6 free parameters, with mild non-linearities coming from the exponential terms in the master equation; Troein *et al.*'s model is a system of 7 ODEs with 19 parameters.

All models were given two time series of $TOC1$ and $CCA1$ protein concentrations sampled hourly across four cycles. The data was obtained by measuring luciferase (LUC) luminescence sampled at regular intervals during 12h:12h light-dark cycles (L:D 12:12) in transgenic lines where LUC was fused to the protein of interest. We indicate these data as translational reporter

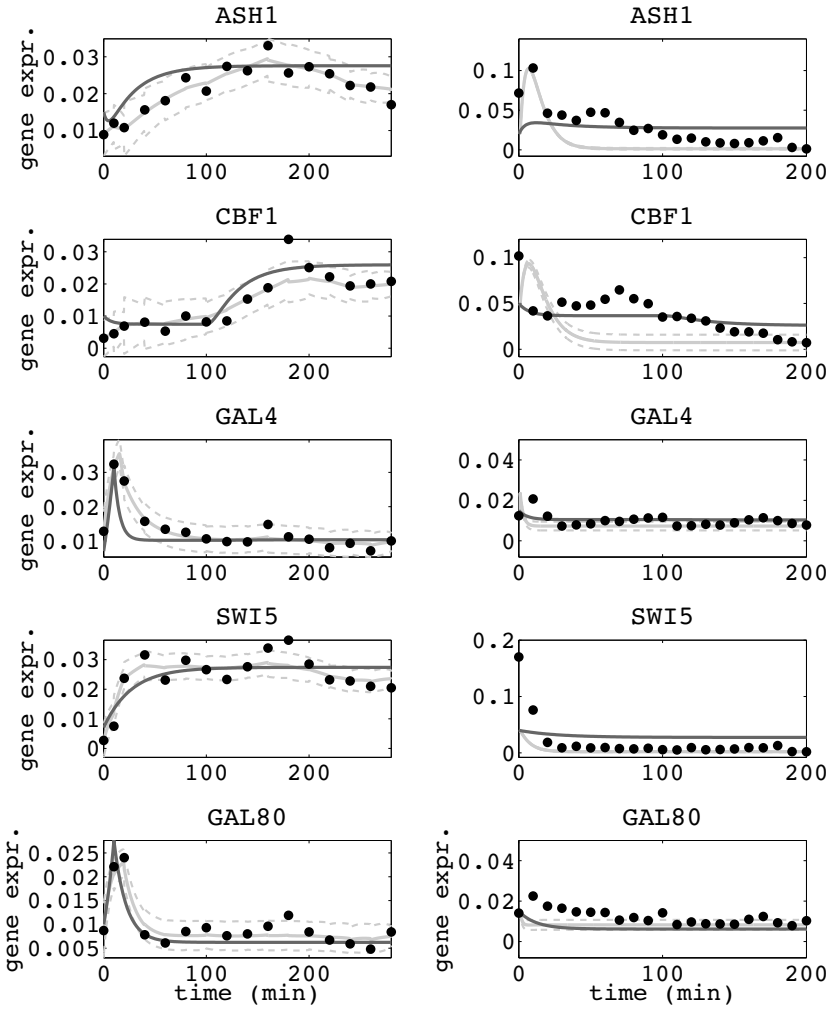

Figure 5.4: Model fitting and predictions on IRMA data set. Left column represents the fit of the models to the training, switch-on data. Right column are predictions on the switch-off data. Results of the Cantone *et al.*'s model (solid dark grey), results of the hybrid model (solid light grey) with confidence intervals (dashed light grey), observed data (black dots).

data (y_{TOC1} and y_{CCA1}). The parameters of the ODE model were determined again by stochastic optimisation, while in our model they are learned during the variational Bayesian scheme as described in Section 5.4.

Figure 5.5 shows the results of applying these procedures on the training data: the top two panels shows the posterior mean of the hybrid regulatory model, while the bottom two panels show the optimised fit of the ODE model. The Troein *et al.*'s model is sufficiently complex to afford a credible fit to the data. Nevertheless, even with its complexity, it cannot explain the higher level of $CCA1$ expression in the third cycle, while the stochastic model of course can accommodate that by a slightly higher activation of the $CCA1$ (latent) promoter variable.

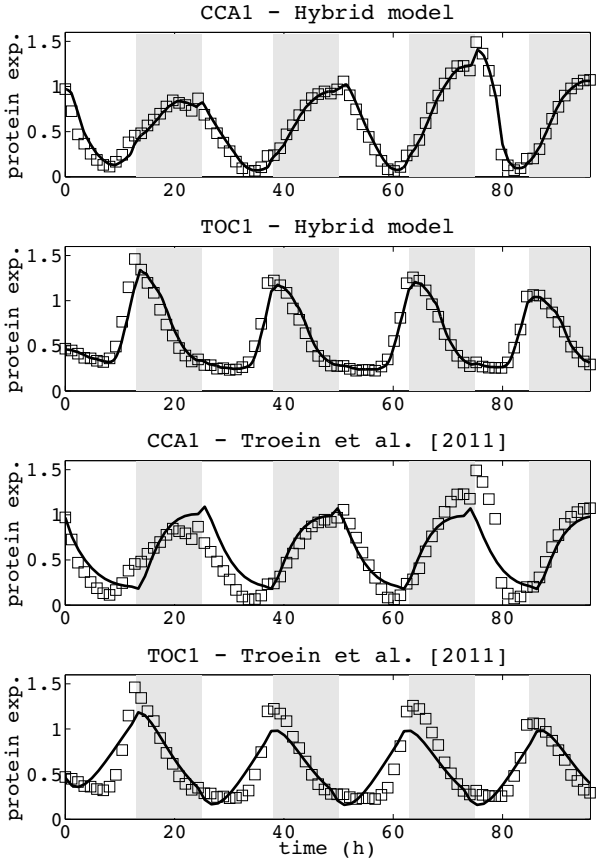

Figure 5.5: Fit to training data: posterior mean of hybrid regulatory model (upper panels); model of (Troein et al., 2011) (bottom panels); observed data (empty squares).

Next, we compare the predictions of the two models on independent data where the light input (photoperiods, i.e. length of the light periods) was altered. We focus here on predicting the expression of $TOC1$ as it is not directly affected by the light input. To do this, we use the two models parameterised using the L:D 12:12 translational reporter data. We then simulate an entraining phase where the oscillator is driven for a long time by a L:D 12:12 cycle and then suddenly alter the photoperiod of the cycle to L:D 6:18, followed by a period of constant light. This mimics the experimental setting in which the data were collected (Troein et al., 2011).

We stress that these predictions are truly *out of sample* predictions in a statistical sense: the data we compare to has not been used in any form to parameterise or tune the models (except for a global scaling factor due to the arbitrariness of the units on the LUC signal). Figure 5.6 shows the results of the simulation of $TOC1$ expression from the two models. Both models accurately predict a reduction of amplitude of the oscillations during the altered L:D cycles. However, Troein *et al.*'s model completely misfits the final constant light period, both in terms of frequency and amplitude. On the other hand, the stochastic hybrid approach provides robust predictions which continue to oscillate with the same period after the change to constant light. Furthermore, it predicts an increase in the average value of the $TOC1$ protein, and a dampening of the amplitudes of the oscillations in constant light.

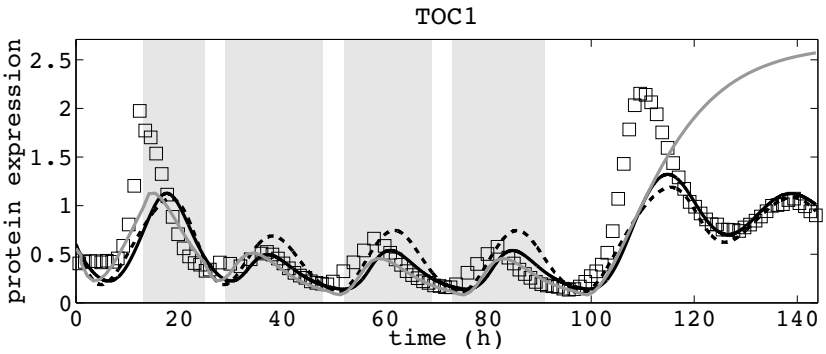

Figure 5.6: Prediction of $TOC1$ protein level after exposure to light dark cycles with altered photoperiod: prediction using the hybrid NFL model (dashed); prediction using the Troein *et al.*'s model (Troein et al., 2011) (grey solid line); prediction using the hybrid repressilator $TOC1$-X-$CCA1$ model (black solid line); observed data (empty squares).

5.7.1 Comparison of the stochastic model with its deterministic version

As a further control, here we consider whether a simple deterministic approach could yield good fits and predictions. We compare the fits and predictions on the *O. tauri* circadian clock data, using two NFL models: our hybrid regulatory model (Eq. 5.9), trained with the variational Bayesian scheme, and its deterministic version, trained by stochastic optimisation (as we have done for Cantone *et al.*'s model and Troein *et al.*'s model).

The deterministic version is a simple ODE model representing the mean behaviour of our hybrid model. It is obtained by replacing the promoter state $\mu(t)$ and protein state $x(t)$, with their mean values. Mathematically, the model is described by the following ODE system:

$$\begin{aligned}
\frac{\mathrm{d}p_{\mu_1}(1,t)}{\mathrm{d}t} &= -k_m\, p_{\mu_1}(1,t) + k_p \exp\left[k_e \langle x_2(t)\rangle\right] p_{\mu_1}(0,t) \\
\frac{\mathrm{d}\langle x_1(t)\rangle}{\mathrm{d}t} &= A_1 p_{\mu_1}(1,t) + b_1 - \lambda_1 \langle x_1(t)\rangle \\
\frac{\mathrm{d}p_{\mu_2}(1,t)}{\mathrm{d}t} &= -k_m\, p_{\mu_2}(1,t) + k_p \exp\left[k_e \langle x_1(t)\rangle\right] p_{\mu_2}(0,t) \\
\frac{\mathrm{d}\langle x_2(t)\rangle}{\mathrm{d}t} &= A_2 p_{\mu_2}(1,t) + b_2 - \lambda_2 \langle x_2(t)\rangle\,,
\end{aligned} \tag{5.32}$$

where we used $p_{\mu_1}(1,t)$ and $p_{\mu_2}(1,t)$ to indicate the marginal probability to be in the active state at time t for the first and second promoters, respectively.

The results show that the simple deterministic model can yield robust predictions (Fig. 5.8), but is unable to fit the training data satisfactorily (Fig. 5.7): the stochastic hybrid regulatory model appears to strike a good compromise between the flexibility given by the latent promoter process, and the robustness given by the simplicity of the underlying model.

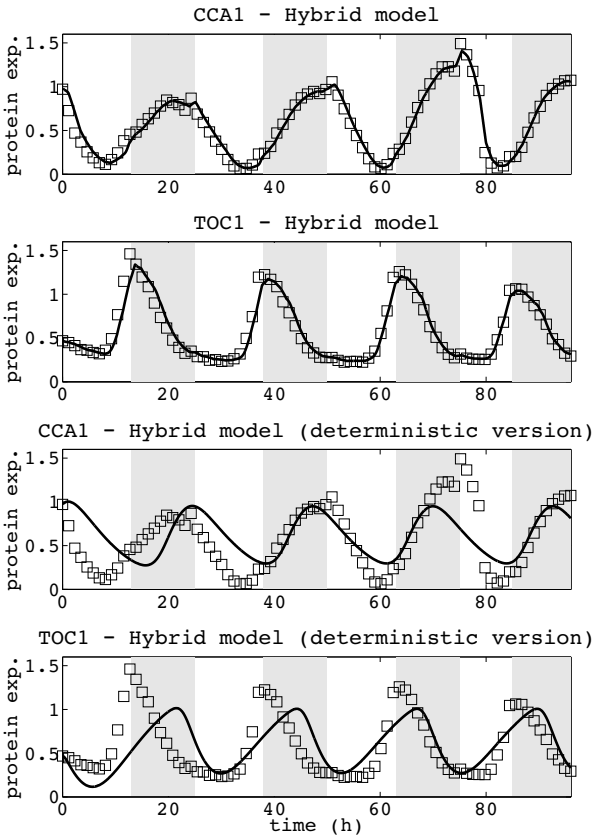

Figure 5.7: Fit to training data: posterior mean of hybrid regulatory model (upper panels); deterministic version of hybrid regulatory model (bottom panels); observed data (empty squares).

Comparing the two models in Figure 5.8, we also note that the predictions of the hybrid regulatory model trained using a Bayesian approach are substantially better than its deterministic version, indicating that the extra stochasticity of the promoter state enabled the model to identify more accurate parameters.

5.7.2 Predicting the clock's structure

Troein *et al.* (Troein et al., 2011) also provide an indirect measurement of promoter states in the form of luciferase time series. This is obtained by inserting in the *Ostreococcus* genome another copy of the $TOC1$ or $CCA1$ promoters directly fused to luciferase. We call these additional data sources transcriptional reporters and denote them as y_{pTOC1} and y_{pCCA1}.

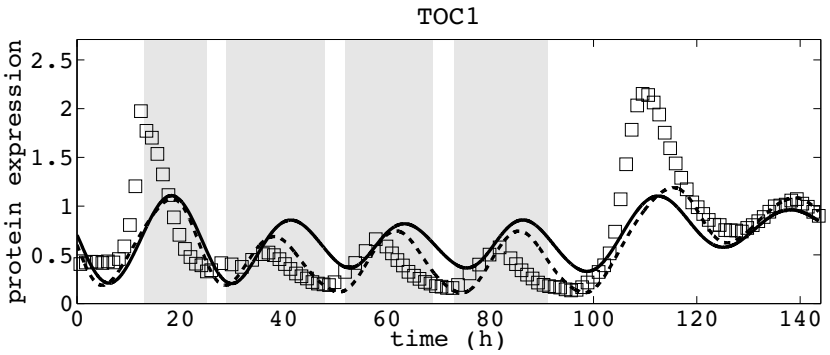

Figure 5.8: Prediction of $TOC1$ protein level after exposure to light dark cycles with altered photoperiod: prediction using the hybrid NFL model (dashed) trained with variational Bayesian scheme; prediction using the deterministic version of the hybrid NFL model (black solid line); observed data (empty squares).

Statistically, these two data types could be represented with the simpler models of Figure 5.9B; nevertheless, obviously the promoter state profiles inferred from transcriptional reporters using the model in Figure 5.9B should match reasonably well the profiles inferred from translational reporters using the model in Figure 5.9A.

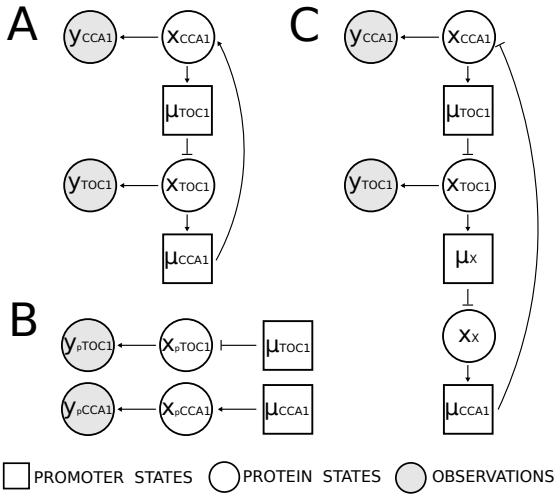

Figure 5.9: Statistical models for *O. tauri*: negative feedback loop (A), transcriptional models (B), repressilator $TOC1$-X-$CCA1$ (C). Note that in order to compare inference results obtained with transcriptional and repressilator models, we need to consider a repressive regulation between μ_{CCA1} and x_{pCCA1}.

Figure 5.10 shows the results of this approach for $CCA1$ (left) and $TOC1$ (right) arranged in a negative feedback loop. Surprisingly, while the predicted promoter states of $TOC1$ match well, $CCA1$ promoters present different dynamics when inferred from transcriptional and translational reporter data, exhibiting an average phase shift of approximately 40^o. More worryingly, the phase shift is highly asymmetrical, with accurately matched off time estimates and widely

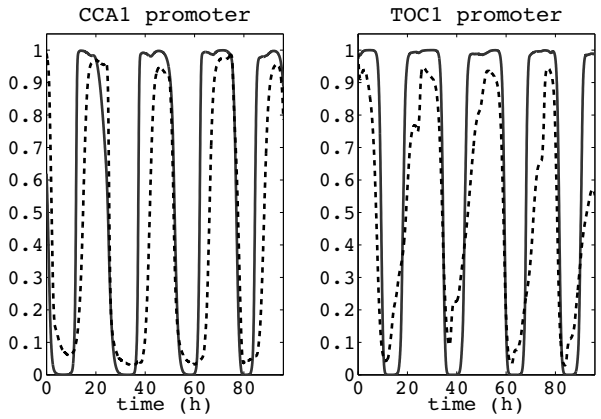

Figure 5.10: Inferred promoter states for $CCA1$ and $TOC1$. Results obtained with the hybrid NFL model (solid lines) using translational reporters; results obtained with the model in Figure 5.9B (dashed lines) using the transcriptional reporters.

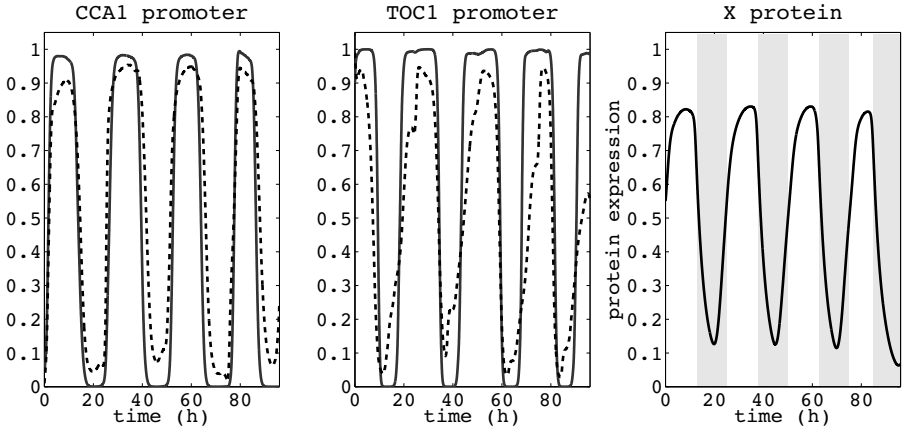

Figure 5.11: Inferred promoter states for $CCA1$ (left) and $TOC1$ (center). Results obtained with the hybrid repressilator model $TOC1$-X-$CCA1$ (solid lines) using translational reporters; results obtained with the model in Figure 5.9B (dashed lines) using transcriptional reporters. Right: mean prediction of the hypothetical gene X.

divergent on time estimates.

We then decided to explore the possibility that this mismatch may be due to an incorrect network topology. Recent results have shown that, in *Arabidopsis thaliana*, $TOC1$ acts as a repressor (Huang et al., 2012), and that the core structure of the *Arabidopsis* clock is better represented as a three nodes network known as a repressilator (Pokhilko et al., 2012; Elowitz and Leibler, 2000). Therefore, we introduced a hypothetical third gene X in the NFL network, leading to a more complex clock network. There are only two possible repressilator configurations after introduction of a putative clock gene X into the previous network, depending on whether X is repressed by $TOC1$ or by $CCA1$. We will refer to them as $TOC1$-X-$CCA1$ and $CCA1$-X-$TOC1$, where X has the role of repressor for $CCA1$ and $TOC1$, respectively (Fig. 5.9C).

We then repeated the inference of the promoter states using the repressilator model. Naturally in this case we do not have translational and transcriptional reporter data for the hypothetical gene X; however, marginalisation of this additional latent variable is straightforward in the Bayesian setting. Therefore, we can use the translational reporter data y_{TOC1} and y_{CCA1} to infer the promoter states μ_{TOC1}, μ_{CCA1} and μ_X (as well as the protein states for all three genes).

Figure 5.11 (left) shows the predicted promoter state of $CCA1$ using the $TOC1$-X-$CCA1$ architecture (the $TOC1$ promoter gives a very good agreement also with this architecture). As can be seen, the average phase shift is greatly reduced, and the inferred promoter states overlap symmetrically.

Interestingly, the $CCA1$-X-$TOC1$ repressilator structure fails to predict the $CCA1$ promoter state (Appendix D.2). Therefore, our approach predicts that the *O. tauri* clock should have a repressilator structure, and that the third gene should be repressed by $TOC1$. As $TOC1$ is expressed mainly in the evening, it follows that the third gene X should be an afternoon gene, as predicted by our model also (Fig. 5.11, right). This is consistent with the existing knowledge of the *A. thaliana* clock.

Next, we checked whether the repressilator model trained on L:D 12:12 translational reporter data is able to predict $TOC1$ profiles in altered photoperiods. The results are shown in Figure 5.6, which compares the repressilator predictions (black solid line) with the NFL predictions (dashed). It is apparent that the repressilator provides a more accurate prediction, particularly during the final constant light period. Further predictions on different altered photoperiods are shown in Appendix D.2.

5.8 Discussion

Mathematical modelling of GRNs is fundamental to our attempts to understand the structure and dynamics of gene networks. Nonlinear ODE models provide an excellent framework to elucidate and predict complex regulatory mechanisms in small to medium scale GRNs. However, they can be vulnerable to incomplete knowledge of the system, and calibrating complex models to limited data may pose an unsurmountable statistical challenge.

Here, we have presented a statistical approach to modelling transcription-translation net-

works which aims at retaining the flexibility allowed by nonlinear ODE models while making possible a statistical exploration of the model's parameterisation. The approach relies on a stochastic hybrid representation of the system where the transcription-translation mechanism is modelled using only two variables: promoter (latent) states and protein states. By replacing complex nonlinearities and additional unknown parameters of ODE models with latent variables, the model becomes simpler, more robust and more identifiable.

Our empirical study demonstrates the identifiability of our approach, and shows how on two real biological problems it can yield comparable or better predictions than competing methods, as well as leading to novel, testable biological hypotheses. Our prediction that a repressilator structure underpins the *O. tauri* clock is in line with recent discoveries on the structure of the *A. thaliana* clock (Pokhilko et al., 2012), and indeed with the structure of known animal circadian clocks (Ukai-Tadenuma et al., 2011). If validated, this finding would suggest that the repressilator structure is an evolutionarily conserved feature of eukaryotic circadian clocks. Furthermore, our model predicts that $TOC1$ acts as a repressor (while remaining an indirect activator of $CCA1$ through a double repression); again, the repressor role of $TOC1$ was recently demonstrated in *A. thaliana* (Huang et al., 2012; Gendron et al., 2012), leading further weight to our hypotheses. While the repressilator model substantially ameliorates the model misfit of the NFL model, there remains some residual unexplained discrepancy between inferences from transcriptional and translational reporters. While this may be due to noise in the data, it cannot be excluded that the complexity of the *O. tauri* oscillator may be even greater, as is the case of other plant oscillators (Pokhilko et al., 2012).

We believe that these results show the promise of this approach as an effective tool in addressing systems biology problems. Nevertheless, this work opens further avenues for development. From the biological point of view, validation of the novel structure of the *O. tauri* clock would be an important step, which is likely to require substantial bioinformatics research. However, perhaps even more interesting would be to computationally explore the links between the transcription-translation oscillator we study and the recently described non-transcriptional oscillator of *O. tauri* (O'Neill et al., 2011). From the computational point of view, this study does not address the important problem of *de novo* reconstruction of the structure of the regulatory network, but relies on pre-existing network structures. A systematic method to combine structure learning with dynamical modelling remains a desirable goal (Oates et al., 2012); we hope that this work will represent an advance in that promising direction.

Chapter 6

Inference from discrete data

As we mentioned in the introduction, a Markov jump process is described in terms of the master equation. Both the deterministic motion and the stochastic fluctuations of the physical system arise naturally by describing the system in terms of the master equation, therefore a jump process makes the description more realistic possible.

In the method we proposed in the previous chapters, we used a chemical master (CME) equation to describe the state of the transcription factor (or the promoter) but a stochastic differential equation (SDE) to describe the state of the mRNA (or the protein). This means that while for the transcription factor we maintain a very satisfactory description, instead for the mRNA we separate the origin of the deterministic and stochastic component. This approximation for the mRNA is valid in a certain limit which depends on the size of the system.

In this chapter we discuss how the method presented in this thesis compare with other methods. In order to do that, in Section 6.1 we consider the simple case of a transcription factor regulating the expression of a single target gene. By treating both transcription factor state μ and mRNA state x as discrete, we can describe the system in terms of the chemical master equation for the joint distribution function $p(\mu, x, t)$. When considering linear reactions, inference can be done by solving exactly the master equation, through equations for the moments. So this represents an excellent case to benchmark how well the method proposed in this thesis works.

In Section 6.2, we discuss the application of the method to nonlinear systems. In this case, the CME cannot be solved exactly for its moments and an approximate solution as the one presented in this thesis becomes useful. In Chapter 5 we have already seen an example of such nonlinear system, where the switching rates of the telegraph process depend on a nonlinear function (i.e. an exponential function) of the protein state. In this section we report a further example, where the protein degrades nonlinearly.

Finally, in Section 6.3 we discuss more in detail the linear noise approximation, as a general method to pass from a description in terms of CME into one in terms of SDE.

6.1 Comparison of the approximate inference method with inference from the chemical master equation

6.1.1 Exact inference from the chemical master equation

Exact inference[1] for gene regulatory networks described in terms of CME, was performed by Boys et al. (Boys et al., 2008).

They develop a fully Bayesian method to estimate the kinetic rate constants (i.e. parameters) of stochastic kinetic systems and apply the method to the Lotka-Volterra model (Lotka, 1910; Volterra, 1926). They first consider the case of complete data, where the entire process is observed in an interval $[0, T]$. By writing a factorised form for the likelihood, which allows to include conjugate priors for the parameters, they derive a posterior distribution (where parameters are independent) and solve easily the parameter estimation problem. In the case of discrete observations, they compute the posterior over the parameters by using the following MCMC scheme: a first block simulate the process, conditional on data and parameters; a second block simulates the parameters, conditional on the process computed in the first block. In particular they describe two different strategies (i.e. a reversible jump and a block updating method) to simulate the entire process through the simulation of the latent process in each interval between discrete observations. Finally they present a way to solve the parameter estimation problem when one of the chemical species is missing. Parameters result identifiable in all cases, including the last worse case when the data is observed only partially.

The article shows that exact inference from discrete data, using the CME, is possible. Inference from the CME becomes relevant when the discreteness of the underlying process cannot be neglected (e.g. chemical species are present in low copy number). However, as the authors point out, the procedure is computationally expensive; therefore an extension to inference in systems larger than the Lotka-Volterra (involving only two species and three reactions) is not straightforward. Another limit of the method is the fact that observations are considered without noise. Inclusion of a noise source in the data would add to the parameter estimation problem also a filtering problem which could be non-trivially solved by using a forward-backward procedure.

An alternative way to infer latent states and estimate reaction rates of a stochastic system can be done by computing the moment of the joint CME, as described below.

6.1.2 Inference using the moments of the chemical master equation

We present a method to do inference by using the moments of the CME. We first derive the equations for the moments, using different procedures. Then we describe how to do state inference and parameter estimation using these equations in a MCMC method. As an example, we apply this method to a stochastic linear system composed of a promoter which regulates

[1] Note that here we use the term "exact inference" to indicate that inference is done directly from the CME. In previous chapters, we used the same term to describe inference in a hybrid continuous-discrete system by using directly the Chapman-Kolmogorov equations, to make distinction with the variational inference approach.

the transcription of a single gene. Finally we present how it is possible to extend the method to nonlinear systems, by using a moment-closure technique.

Derivation of ODEs for the evolution of the moments

To derive ODEs for the moments we consider the simple Lotka-Volterra reaction system; the same procedure is then easily applied to the promoter-mRNA system. The Lotka-Volterra system is composed of two species (n_1 and n_2) whose stochastic dynamics is described by the following three reactions:

$$R_1: \quad n_1 \xrightarrow{c_1} 2n_1 \qquad R_2: \quad n_1 + n_2 \xrightarrow{c_2} 2n_2 \qquad R_3: \quad n_2 \xrightarrow{c_3} 0$$

from which we can write the following transition probabilities for an infinitesimal time Δt

$$\begin{aligned}
&\mathrm{Prob}(n_1 \to n_1 + 1;\ n_2 \to n_2) = c_1 n_1 \Delta t \\
&\mathrm{Prob}(n_1 \to n_1 - 1;\ n_2 \to n_2 + 1) = c_2 n_1 n_2 \Delta t \\
&\mathrm{Prob}(n_1 \to n_1;\ n_2 \to n_2 - 1) = c_3 n_2 \Delta t \\
&\mathrm{Prob}(n_1 \to n_1;\ n_2 \to n_2) = 1 - (c_1 n_1 + c_2 n_1 n_2 + c_3 n_2)\Delta t
\end{aligned}$$

From these we finally obtain the joint CME for all the molecular species

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} p(n_1, n_2) &= c_1(n_1 - 1)p(n_1 - 1, n_2) + c_2(n_1 + 1)(n_2 - 1)p(n_1 + 1, n_2 - 1) \\
&+ c_3(n_2 + 1)(n_1, n_2 + 1) - (c_1 n_1 + c_2 n_1 n_2 + c_3 n_2)p(n_1, n_2)
\end{aligned}$$

where we have simplified the notation by removing the time variable. The ODE for the first moment for molecular species $\langle n_1 \rangle = \sum_{n_1} n_1 p(n_1, n_2)$ is given by

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \sum_{n_1} n_1 p(n_1, n_2) &= \sum_{n_1} \Big[n_1 c_1 (n_1 - 1) p(n_1 - 1, n_2) \\
&+ n_1 c_2 (n_1 + 1)(n_2 - 1) p(n_1 + 1, n_2 - 1) + n_1 c_3 (n_2 + 1)(n_1, n_2 + 1) \\
&- n_1 (c_1 n_1 + c_2 n_1 n_2 + c_3 n_2) p(n_1, n_2) \Big]
\end{aligned}$$

which we can rewrite as[2]

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \sum_{n_1} n_1 p(n_1, n_2) &= \sum_{n_1} \Big[(n_1 + 1) c_1 n_1 p(n_1, n_2) \\
&+ (n_1 - 1) c_2 n_1 n_2 p(n_1, n_2) + n_1 c_3 n_2 (n_1, n_2) \\
&- n_1 (c_1 n_1 + c_2 n_1 n_2 + c_3 n_2) p(n_1, n_2) \Big] .
\end{aligned}$$

[2]This is simply obtained by changing the variables, e.g. by substituting n_1 with $n_1 + 1$ we obtain $\sum_{n_1} n_1 c_1 (n_1 - 1) p(n_1 - 1, n_2) = \sum_{n_1} (n_1 + 1) c_1 n_1 p(n_1, n_2)$. This will not affect the result, as the summation occurs into the interval $[0, +\infty]$.

and finally as

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle n_1\rangle = c_1\langle n_1\rangle - c_2\langle n_1 n_2\rangle = c_1\langle n_1\rangle - c_2\mathrm{cov}(n_1,n_2) - c_2\langle n_1\rangle\langle n_2\rangle\,.$$

In the last equivalence we have simply used the relation for the covariance $\mathrm{cov}(n_1,n_2) = \langle n_1 n_2\rangle - \langle n_1\rangle\langle n_2\rangle$.

By following this procedure, we can obtain ODEs directly from the joint CME for all first, second and higher-order moments.

Alternatively, ODEs for the moments can be obtained by means of the moment generating function. For a stochastic system with two molecular species, the moment generating function G is (Gardiner, p. 273)

$$G(s_{n_1}, s_{n_2}) \quad = \quad \sum_{s_{n_1}}\sum_{s_{n_2}} \left[s_{n_1}^{n_1} s_{n_2}^{n_2} p(n_1,n_2)\right]\,,$$

where we have again omitted the time variable. By definition, all moments can be computed from the moment generating function. For example, through the following partial derivations we get the first and second moment of n_1 and a cross-moment:

$$\begin{aligned}
\frac{\partial G}{\partial s_{n_1}} &= \sum_{s_{n_1}}\sum_{s_{n_2}} n_1 s_{n_1}^{n_1-1} s_{n_2} p(n_1,n_2)\Big|_{\substack{s_{n_1}=1\\ s_{n_2}=1}} &= \langle n_1\rangle \\
\frac{\partial G^2}{\partial s_{n_1}^2} &= \sum_{s_{n_1}}\sum_{s_{n_2}} n_1(n_1-1) s_{n_1}^{n_1-2} s_{n_2}^{n_2} p(n_1,n_2)\Big|_{\substack{s_{n_1}=1\\ s_{n_2}=1}} &= \left\langle n_1^2\right\rangle - \langle n_1\rangle \\
\frac{\partial G^2}{\partial s_{n_1}\partial s_{n_2}} &= \sum_{s_{n_1}}\sum_{s_{n_2}} n_1 s_{n_1}^{n_1-1} n_2 s_{n_2}^{n_2-1} p(n_1,n_2)\Big|_{\substack{s_{n_1}=1\\ s_{n_2}=1}} &= \langle n_1 n_2\rangle\,.
\end{aligned}$$

In order to obtain ODE for these moments, we need to do the derivative of the function G with respect to time. Using the change of the variables as before, we get

$$\begin{aligned}
\frac{\partial G}{\partial t} &= \sum_{s_{n_1}}\sum_{s_{n_2}} \Big[s_{n_1}^{n_1+1} s_{n_2}^{n_2} c_1 n_1 + s_{n_1}^{n_1-1} s_{n_2}^{n_2+1} c_2 n_1 n_2 + s_{n_1}^{n_1} s_{n_2}^{n_2-1} c_3 n_2 \\
&- s_{n_1}^{n_1} s_{n_2}^{n_2} c_1 n_1 - s_{n_1}^{n_1} s_{n_2}^{n_2} c_2 n_1 n_2 - s_{n_1}^{n_1} s_{n_2}^{n_2} c_3 n_2\Big] p(n_1,n_2) \\
&= \sum_{s_{n_1}}\sum_{s_{n_2}} \Big[s_{n_1}^{n_1-1} s_{n_1}^2 s_{n_2}^{n_2} c_1 n_1 + s_{n_1}^{n_1-1} s_{n_2}^{n_2-1} s_{n_2}^2 c_2 n_1 n_2 + s_{n_1}^{n_1} s_{n_2}^{n_2-1} c_3 n_2 \\
&- s_{n_1}^{n_1-1} s_{n_1} s_{n_2}^{n_2} c_1 n_1 - s_{n_1}^{n_1-1} s_{n_1} s_{n_2}^{n_2-1} s_{n_2} c_2 n_1 n_2 - s_{n_1}^{n_1} s_{n_2}^{n_2-1} s_{n_2} c_3 n_2\Big] p(n_1,n_2) \\
&= \sum_{s_{n_1}}\sum_{s_{n_2}} \Big[\left(n_1 s_{n_1}^{n_1-1} s_{n_2}^{n_2}\right)\left(s_{n_1}^2 - s_{n_1}\right) c_1 + \left(n_1 s_{n_1}^{n_1-1} n_2 s_{n_2}^{n_2-1}\right)\left(s_{n_2}^2 - s_{n_1} s_{n_2}\right) c_2 \\
&+ \left(n_2 s_{n_2}^{n_2-1} s_{n_1}^{n_1}\right)(1 - s_{n_2})\, c_3\Big] p(n_1,n_2)
\end{aligned}$$

which we can rewrite in the following form

$$\frac{\partial G}{\partial t} = c_1 s_{n_1}(s_{n_1}-1)\frac{\partial G}{\partial s_{n_1}} + c_2 s_{n_2}(s_{n_2}-s_{n_1})\frac{\partial^2 G}{\partial s_{n_1}\partial s_{n_2}} + c_3(1-s_{n_2})\frac{\partial G}{\partial s_{n_2}}\,.$$

Then, the ODE for the first moment of n_1 can be derived as

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\langle n_1\rangle &= \frac{\partial}{\partial s_{n_1}}\left(\frac{\partial G}{\partial t}\right)\Bigg|_{\substack{s_{n_1}=1\\ s_{n_2}=1}} \\
&= c_1 s_{n_1}(s_{n_1}-1)\frac{\partial^2 G}{\partial s_{n_1}^2} + c_1(2s_{n_1}-1)\frac{\partial G}{\partial s_{n_1}} + c_2 s_{n_2}(s_{n_2}-s_{n_1})\frac{\partial^3 G}{\partial s_{n_1}^2 \partial s_{n_2}} \\
&+ c_2(-s_{n_2})\frac{\partial^2 G}{\partial s_{n_1}\partial s_{n_2}} + c_3(1-s_{n_2})\frac{\partial^2 G}{\partial s_{n_1}\partial s_{n_2}} \\
&= c_1\frac{\partial G}{\partial s_{n_1}} - c_2\frac{\partial^2 G}{\partial s_{n_1}\partial s_{n_2}} = c_1\langle n_1\rangle - c_2\langle n_1 n_2\rangle\ .
\end{aligned}$$

The ODE for the second moment of n_1 can be derived in a similar way. We start from the following partial derivative

$$\begin{aligned}
\frac{\partial^2}{\partial s_{n_1}^2}\left(\frac{\partial G}{\partial t}\right) &= \left[\frac{\mathrm{d}}{\mathrm{d}t}\left\langle n_1^2\right\rangle - \frac{\mathrm{d}}{\mathrm{d}t}\langle n_1\rangle\right] \\
&= \frac{\partial}{\partial s_{n_1}}\left[c_1 s_{n_1}(s_{n_1}-1)\frac{\partial^2 G}{\partial s_{n_1}^2} + c_1(2s_{n_1}-1)\frac{\partial G}{\partial s_{n_1}} + c_2 s_{n_2}(s_{n_2}-s_{n_1})\frac{\partial^3 G}{\partial s_{n_1}^2\partial s_{n_2}}\right. \\
&- \left. c_2 s_{n_2}\frac{\partial^2 G}{\partial s_{n_1}\partial s_{n_2}} + c_3(1-s_{n_2})\frac{\partial^2 G}{\partial s_{n_1}\partial s_{n_2}}\right] \\
&= c_1(2s_{n_1}-1)\frac{\partial^2 G}{\partial s_{n_1}^2} + c_1 s_{n_1}(s_{n_1}-1)\frac{\partial^3 G}{\partial s_{n_1}^3} + 2c_1\frac{\partial G}{\partial s_{n_1}} + c_1(2s_{n_1}-1)\frac{\partial^2 G}{\partial s_{n_1}^2} \\
&- c_2 s_{n_2}\frac{\partial^3 G}{\partial s_{n_1}^2\partial s_{n_2}} + c_2 s_{n_2}(s_{n_2}-s_{n_1})\frac{\partial^4 G}{\partial s_{n_1}^3\partial s_{n_2}} - c_2 s_{n_2}\frac{\partial^3 G}{\partial s_{n_1}^2\partial s_{n_2}} + c_3(1-s_{n_2})\frac{\partial^3 G}{\partial s_{n_1}^2\partial s_{n_2}}\,,
\end{aligned}$$

which, by setting $s_{n_1} = s_{n_2} = 1$, becomes

$$\frac{\partial^2}{\partial s_{n_1}^2}\left(\frac{\partial G}{\partial t}\right) = \left[\frac{\mathrm{d}}{\mathrm{d}t}\left\langle n_1^2\right\rangle - \frac{\mathrm{d}}{\mathrm{d}t}\langle n_1\rangle\right] = 2c_1\left(\frac{\partial^2 G}{\partial s_{n_1}^2} + \frac{\partial G}{\partial s_{n_1}}\right) - 2c_2\frac{\partial^3 G}{\partial s_{n_1}^2\partial s_{n_2}}\,.$$

The ODE for the second moment is then

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\left\langle n_1^2\right\rangle &= \left[\frac{\partial^2}{\partial s_{n_1}^2}\left(\frac{\partial G}{\partial t}\right) + \frac{\mathrm{d}}{\mathrm{d}t}\langle n_1\rangle\right] = 2c_1\left(\frac{\partial^2 G}{\partial s_{n_1}^2} + \frac{\partial G}{\partial s_{n_1}}\right) - 2c_2\frac{\partial^3 G}{\partial s_{n_1}^2\partial s_{n_2}} + c_1\langle n_1\rangle - c_2\langle n_1 n_2\rangle \\
&= 2c_1\left[\left\langle n_1^2\right\rangle - \langle n_1\rangle + \langle n_1\rangle\right] - 2c_2\left[\left\langle n_1^2 n_2\right\rangle - \langle n_1 n_2\rangle\right] + c_1\langle n_1\rangle - c_2\langle n_1 n_2\rangle \\
&= 2c_1\left\langle n_1^2\right\rangle + c_1\langle n_1\rangle - 2c_2\left\langle n_1^2 n_2\right\rangle + c_2\langle n_1 n_2\rangle\ ,
\end{aligned}$$

where we have used the following partial derivative

$$\frac{\partial^3 G}{\partial s_{n_1}^2\partial s_{n_2}} = \left\langle n_1^2 n_2\right\rangle - \langle n_1 n_2\rangle\ .$$

We conclude this subsection with a third more efficient and elegant way to obtain ODEs for the moments (Grima, 2012). We first write down the stoichiometric matrix S and propensity function $\hat{f}$ for the Lotka-Volterra system:

$$S \equiv \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \qquad \hat{f} \equiv \left(c_1\frac{n_1}{\Omega},\ c_2\frac{n_1 n_2}{\Omega^2},\ c_3\frac{n_2}{\Omega}\right)\,, \tag{6.1}$$

where rows and columns in the stoichiometric matrix represent molecular species (n_1, n_2) and reactions (R_1, R_2, R_3) respectively. The quantity Ω represents the volume of the system. ODEs for the evolution of the moments can be obtained by using general formulas derived directly from the CME. For example, first and second moments can be simply obtained by using the following formulas:

$$\begin{aligned}
\frac{\partial}{\partial t}\frac{\langle n_i\rangle}{\Omega} &= \sum_{j=1}^{R} S_{ij}\langle \hat{f}_j\rangle \\
\frac{\partial}{\partial t}\frac{\langle n_i n_k\rangle}{\Omega^2} &= \Omega^{-1}\sum_{j=1}^{R}\left(S_{kj}\langle n_i \hat{f}_j\rangle + S_{ij}\langle n_k \hat{f}_j\rangle + S_{ij}S_{kj}\langle \hat{f}_j\rangle\right),
\end{aligned}$$

where R is the total number of reactions, S_{ij} is the (ij)-th element of the stoichiometric matrix and $\hat{f}_j$ the j-th element of the propensity function.

Equations for the moments in a promoter-mRNA system

Here we derive ODEs for the evolution of the moments in a stochastic system composed of a promoter which regulate the transcription of a mRNA. The promoter is assumed to be switching between two states (active and inactive), therefore we use two variables to represent its dynamics, n_1 for the inactive state and n_2 for the active one, and a single variable n_3 to represent the mRNA[3]. Mathematically, we consider the following chemical reactions:

$$R_1:\ n_1 \xrightarrow{c_1} n_2 \qquad R_2:\ n_2 \xrightarrow{c_2} n_1 \qquad R_3:\ n_2 \xrightarrow{c_3} n_2 + n_3 \qquad R_4:\ n_3 \xrightarrow{c_4} 0$$

from which we easily deduce the following stoichiometric matrix and propensity function:

$$S \equiv \begin{pmatrix} -1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \qquad \hat{f} \equiv \left(c_1\frac{n_1}{\Omega},\ c_2\frac{n_2}{\Omega},\ c_3\frac{n_2}{\Omega},\ c_4\frac{n_3}{\Omega}\right).$$

As described above, stoichiometric matrix and propensity function can be used to derive ODEs for first and second moments

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\langle n_1\rangle &= -c_1\langle n_1\rangle + c_2\langle n_2\rangle \\
\frac{\mathrm{d}}{\mathrm{d}t}\langle n_2\rangle &= -c_2\langle n_2\rangle + c_1\langle n_1\rangle \\
\frac{\mathrm{d}}{\mathrm{d}t}\langle n_3\rangle &= -c_4\langle n_3\rangle + c_3\langle n_2\rangle \\
\frac{\mathrm{d}}{\mathrm{d}t}\langle n_1^2\rangle &= -2c_1\langle n_1^2\rangle + c_1\langle n_1\rangle + 2c_2\langle n_1 n_2\rangle + c_2\langle n_2\rangle \\
\frac{\mathrm{d}}{\mathrm{d}t}\langle n_2^2\rangle &= -2c_2\langle n_2^2\rangle + c_2\langle n_2\rangle + 2c_1\langle n_1 n_2\rangle + c_1\langle n_1\rangle \\
\frac{\mathrm{d}}{\mathrm{d}t}\langle n_3^2\rangle &= -2c_4\langle n_3^2\rangle + c_4\langle n_3\rangle + 2c_3\langle n_2 n_3\rangle + c_3\langle n_2\rangle,
\end{aligned}$$

[3]In particular n_1 and n_2 are binary variables, with $n_1 + n_2 = 1$.

and for the following cross-moments

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\langle n_1 n_2\rangle &= -(c_1+c_2)\langle n_1 n_2\rangle + c_1\langle n_1^2\rangle + c_2\langle n_2^2\rangle - c_1\langle n_1\rangle - c_2\langle n_2\rangle \\
\frac{\mathrm{d}}{\mathrm{d}t}\langle n_2 n_3\rangle &= -(c_2+c_4)\langle n_2 n_3\rangle + c_1\langle n_1 n_3\rangle + c_3\langle n_2^2\rangle \\
\frac{\mathrm{d}}{\mathrm{d}t}\langle n_1 n_3\rangle &= -(c_1+c_4)\langle n_1 n_3\rangle + c_2\langle n_2 n_3\rangle + c_3\langle n_1 n_2\rangle\,.
\end{aligned}$$

When the system is linear as in this case, we can solve exactly the equations for the moments. However, doing inference in this system is non-trivial since an analytical form for the probability density function is not available. A possible way to do inference is by approximating the likelihood with a multivariate Gaussian where mean and covariance are given by the equations for the moments (Milner et al., 2013). This could be object of strong criticism, as it means to approximate some binary variables with a Gaussian distribution. A better solution to the inference problem is provided by (Opper and Sanguinetti, 2008), as we mentioned in Section 1.3.1.

Inference and parameter estimation using ODEs for the moments

Here we describe a way to do inference and parameter estimation using the equations for the moments. We first consider the case when all molecular species are observed at discrete times and we are interested only in the estimation of the chemical reaction rates. This means that in our promoter-mRNA system we have observations for n_1, n_2 and n_3 and we are interested in the estimation of rates c_1, c_2, c_3 and c_4. Figure 6.2 shows a realisation obtained with the Gillespie algorithm (Gillespie, 1977) from the promoter-mRNA system.

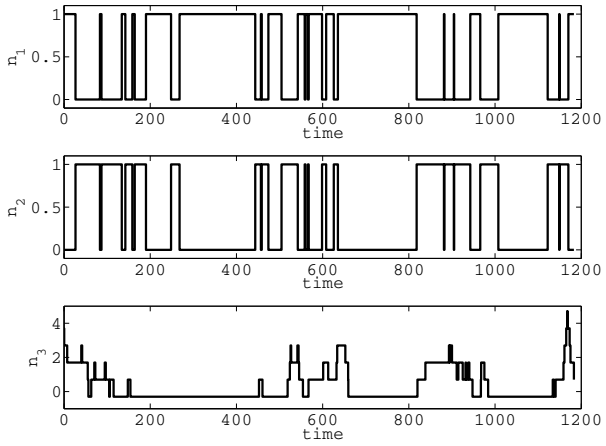

Figure 6.1: Stochastic realisation of the promoter-mRNA system, using the following rates $c_1 = 0.03$, $c_2 = 0.03$, $c_3 = 0.07$, $c_4 = 0.04$ and the following initial conditions $n_1(0) = 1$, $n_2(0) = 0$, $n_3(0) = 3.7$.

A Metropolis MCMC method can be used to solve the parameter estimation problem. Parameter values can be sampled from a multivariate Gaussian distribution. Samples are ac-

cepted/rejected by assuming a Gaussian likelihood for the observed data $\boldsymbol{y}$:

$$\mathcal{L} = \prod_{i=1}^{3} \prod_{k=1}^{N} \mathcal{N}\left(y_i(k) | \langle n_i(t_k) \rangle, \langle n_i(t_k)^2 \rangle - \langle n_i(t_k) \rangle^2\right)$$

where mean $\langle n_i(t)^2 \rangle$ and variance $\langle n_i(t)^2 \rangle - \langle n_i(t) \rangle^2$ for all species are computed by solving the moments of the CME. Therefore, at every iteration of the algorithm, ODEs for the moments have to be simulated in order to evaluate the likelihood.

When data are observed only partially, e.g. mRNA levels are observed but promoter states are latent, parameters estimation is still possible. We follow Milner and colleagues (Milner et al., 2013) and update the latent states through a Metropolis-Hastings step. This step is alternated with the Metropolis step described above to evaluate the likelihood with respect to the parameters.

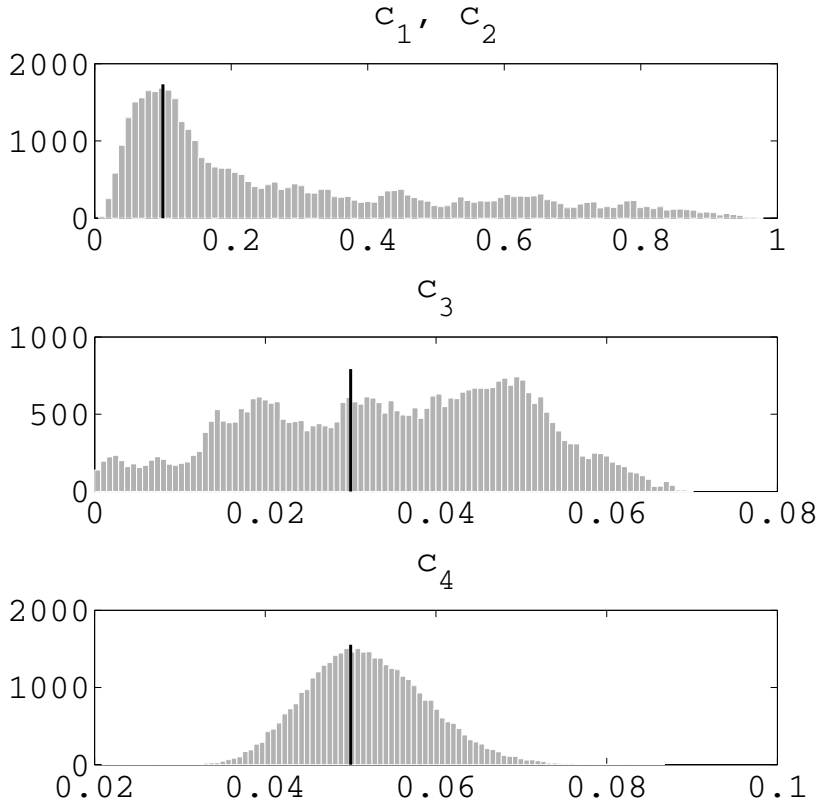

Figure 6.2: Histograms showing samples for each reaction rate parameter (black vertical lines represent real values). The switching system was simulated using the following reaction parameters: $c_1 = c_2 = 0.1$, $c_3 = 0.03$, $c_4 = 0.05$. Variances for the Gaussian proposal of these parameters were: 0.006^2, 0.006^2, 0.0005^2, 0.004^2. Estimated parameters giving maximum likelihood are the following: $c_1 = c_2 = 0.116$, $c_3 = 0.034$, $c_4 = 0.049$.

Moment-closure method

The method described by Boys and colleagues (Boys et al., 2008) can be adopted for stochastic nonlinear systems, such as Lokta-Volterra. On the other hand, the inference method based on the moments, as we described above it is applicable only to linear systems. In fact it is possible to obtain a closed system to solve the ODEs for the moments only if a system is linear. However, by adopting a moment-closure technique (Lee et al., 2009), we can disregard of higher-order moments in nonlinear systems and keep using the same inference method (Milner et al., 2013).

Here we report an example on the promoter-mRNA system we described before, but where now the mRNA degrades quadratically (i.e. nonlinearly). The stochastic system is described by the following chemical reactions:

$$R_1: \ n_1 \xrightarrow{c_1} n_2 \qquad R_2: \ n_2 \xrightarrow{c_2} n_1 \qquad R_3: \ n_2 \xrightarrow{c_3} n_2 + n_3 \qquad R_4: \ n_3 + n_3 \xrightarrow{c_4} 0$$

and the stoichiometric matrix and propensity function become:

$$S \equiv \begin{pmatrix} -1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -2 \end{pmatrix} \qquad \hat{f} \equiv \left(c_1 \frac{n_1}{\Omega},\ c_2 \frac{n_2}{\Omega},\ c_3 \frac{n_2}{\Omega},\ \frac{c_4}{2} \frac{n_3(n_3-1)}{\Omega} \right).$$

ODEs for the evolution of the moments result as the ones in the linear case, except those involving the mRNA species n_3:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\langle n_3 \rangle &= -c_4 \left(\langle n_3^2 \rangle - \langle n_3 \rangle \right) + c_3 \langle n_2 \rangle \\
\frac{\mathrm{d}}{\mathrm{d}t}\langle n_3^2 \rangle &= 2c_3 \langle n_2 n_3 \rangle + c_3 \langle n_2 \rangle + 2c_4 \left(2\langle n_3^2 \rangle - \langle n_3^3 \rangle - \langle n_3 \rangle \right) \\
\frac{\mathrm{d}}{\mathrm{d}t}\langle n_2 n_3 \rangle &= (c_4 - c_2)\langle n_2 n_3 \rangle - c_4 \langle n_2 n_3^2 \rangle + c_1 \langle n_1 n_3 \rangle + c_3 \langle n_2^2 \rangle \\
\frac{\mathrm{d}}{\mathrm{d}t}\langle n_1 n_3 \rangle &= (c_4 - c_1)\langle n_1 n_3 \rangle - c_4 \langle n_1 n_3^2 \rangle + c_2 \langle n_2 n_3 \rangle + c_3 \langle n_1 n_2 \rangle .
\end{aligned}$$

Both the ODE for second moment $\langle n_3^2 \rangle$ and ODEs for cross-moments $\langle n_2 n_3 \rangle$ and $\langle n_1 n_3 \rangle$ depend on higher-order moments. A moment-closure technique allows to disregard of this dependency by closing the moment equations. This can be done with different distributional assumption to close the moments. Here we assume a Gaussian approximation, which means that cumulants of third and higher-order are zero. Third-order cumulants are defined by the following quantity (Gardiner, 2009):

$$\langle\langle xyz \rangle\rangle = \langle xyz \rangle - \langle xy \rangle \langle z \rangle - \langle x \rangle \langle yz \rangle - \langle xz \rangle \langle y \rangle + 2\langle x \rangle \langle y \rangle \langle z \rangle$$

where x_1, x_2, x_3 are three distinct random variables. From this definition we can easily obtain the expression of other third-order cumulants, e.g. when $z = x$ and $z = y = x$ we obtain the

following third-order cumulants, respectively

$$\begin{aligned}\langle\langle x^2y\rangle\rangle &= \langle x^2y\rangle - 2\langle x\rangle\langle xy\rangle - \langle x^2\rangle\langle y\rangle + 2\langle x\rangle^2\langle y\rangle \\ \langle\langle x^3\rangle\rangle &= \langle x^3\rangle - 3\langle x\rangle\langle x^2\rangle + 2\langle x\rangle\langle x\rangle^2 .\end{aligned}$$

By setting third-order cumulants to zero, we obtain the desired closure by having a dependency of third moments on lower order moments. Therefore, we can substitute the third-order moments in our set of ODEs with the following relations:

$$\begin{aligned}\langle n_3^3\rangle &= 3\langle n_3\rangle\langle n_3^2\rangle - 2\langle n_3\rangle\langle n_3\rangle^2 \\ \langle n_1n_3^2\rangle &= 2\langle n_3\rangle\langle n_1n_3\rangle + \langle n_3^2\rangle\langle n_1\rangle - 2\langle n_3\rangle^2\langle n_1\rangle \\ \langle n_2n_3^2\rangle &= 2\langle n_3\rangle\langle n_2n_3\rangle + \langle n_3^2\rangle\langle n_2\rangle - 2\langle n_3\rangle^2\langle n_2\rangle .\end{aligned}$$

By closing the moments in this way, we are assuming a Gaussian approximation from the third and higher-order moments. On the other hand, the variational mean field approximation we described in Chapter 5 we are assuming a Gaussian approximation for all the moments. Other methods which do not rely on particular distributional assumption have also been recently developed (Grima, 2010).

6.2 Application to nonlinear systems

The variational method presented in this thesis can be applied to nonlinear system as well. Here we report a simple example where the protein degrades nonlinearly, by using a quadratic function. We consider the following system, composed of a transcription factor whose binary activity is $\mu(t)$, which regulates a target gene with expression $x(t)$:

$$\begin{aligned}\mu(t) &\sim \mathcal{TP}(f_\pm) \\ \mathrm{d}x &= (A\mu + b - \lambda x^2)\mathrm{d}t + \sigma \mathrm{d}w(t)\end{aligned} \tag{6.2}$$

We follow (Opper et al., 2010) and use the following factorised approximating distribution

$$q(\chi,\nu) = q_\chi(\chi)\, q_\nu(\nu) \tag{6.3}$$

where we defined χ and ν as continuous-time sample paths of the processes $x(t)$ and $\mu(t)$ over a time interval $[0,T]$. The processes $q_\chi(\chi)$ and $q_\nu(\nu)$ represent pure Gaussian diffusion processes and pure telegraph processes, respectively. Plugging the approximating distribution into the KL divergence formula we obtain

$$\begin{aligned}KL[q(\chi,\nu)\|p(\chi,\nu|D,\Theta)] &= \log Z + \langle KL\left[q_\chi(\chi)\|p(\chi|\nu)\right]\rangle_{q_\nu} \\ &- \sum_{i=1}^{N}\langle\log p(y_i|x(t_i))\rangle_{q_\chi} + KL\left[q_\nu(\nu)\|p(\nu)\right] .\end{aligned} \tag{6.4}$$

where $D = [y_1, \ldots, y_N]$ represent the discrete protein observations with i.i.d. Gaussian noise and $\Theta = [A,\, b,\, \lambda]$ are the kinetic model parameters. As in Section 5.4, we report how each term is computed.

1. The first term is constant, therefore is neglected during the minimisation.

2. The second term contains the KL divergence between a Gaussian diffusion process and a nonlinear process. This has been computed in a general form in (Archambeau et al., 2007) and is given by

$$KL\left[q \| p_{sde}\right] = \frac{1}{2}\int_0^T \left\langle (\boldsymbol{d} - \boldsymbol{d}_L)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{d} - \boldsymbol{d}_L) \right\rangle_{q_t} \mathrm{d}t\,, \tag{6.5}$$

where q and p_{sde} are an approximating Gaussian process (with linear drift $\mathbf{d}_L$) and a general process (with drift $\mathbf{d}$), respectively, and $\boldsymbol{\Sigma}$ is the covariance matrix for the noise driving process. The expectation inside the integral is computed with respect to the single time marginal distribution q_t.

In our particular case, $\mathbf{d} \rightarrow A\mu + b - \lambda x^2$ and $\mathbf{d}_L \rightarrow \alpha(t)x + \beta(t)$ represent the prior drift (conditioned on the upstream promoter state) and the approximating linear function, respectively. $\boldsymbol{\Sigma} \rightarrow \sigma^2$ represents the variance of the noise driving process and $q_t \rightarrow q_\chi$ the single time marginal of our approximating Gaussian process.

By solving the argument inside the integral in Equation 6.5 and computing the expectation of the KL divergence with respect to q_ν, we obtain the following expression for the second term:

$$\begin{aligned}
\langle KL\left[q_\chi(\chi) \| p(\chi|\nu)\right]\rangle_{q_\nu} &= \frac{1}{2\sigma^2}\int_0^T \mathrm{d}t \Big[\lambda^2 \left\langle x^4(t)\right\rangle_{q_\chi} + 2\lambda\alpha(t)\left\langle x^3(t)\right\rangle_{q_\chi} \\
&+ \left(\alpha^2(t) - 2A\lambda q_\mu(1,t) - 2b\lambda + 2\beta(t)\lambda\right)\left\langle x^2(t)\right\rangle_{q_\chi} \\
&+ \left(2\alpha(t)\beta(t) - 2A\alpha(t)q_\mu(1,t) - 2b\alpha(t)\right)\langle x(t)\rangle_{q_\chi} \\
&+ (\beta(t) - b)^2 + A^2 q_\mu(1,t) - 2A(\beta(t) - b)q_\mu(1,t)\Big]\,,
\end{aligned} \tag{6.6}$$

where $q_\mu(1,t) = \langle \mu(t)\rangle_{q_\nu}$ and the functions $\langle x(t)\rangle_{q_\chi}$, $\left\langle x^2(t)\right\rangle_{q_\chi}$, $\left\langle x^3(t)\right\rangle_{q_\chi}$ and $\left\langle x^4(t)\right\rangle_{q_\chi}$ represent the first four marginal moments of the approximating process $q_\chi(\chi)$. Under the assumption of a Gaussian process approximation, these moments are simply given by

$$\begin{aligned}
\langle x(t)\rangle_{q_\chi} &= m(t) \\
\left\langle x^2(t)\right\rangle_{q_\chi} &= m^2(t) + c^2(t) \\
\left\langle x^3(t)\right\rangle_{q_\chi} &= m^3(t) + 3m(t)c^2(t) \\
\left\langle x^4(t)\right\rangle_{q_\chi} &= m^4(t) + 6m^2(t)c^2(t) + 3c^4(t)
\end{aligned}$$

where we have indicated with $c^2(t)$ the single time variance of the approximating diffusion process $q_\chi(\chi)$.

3. The third term contains the likelihood part (see Equation 5.16).

4. The fourth and last term is computed using the variational approximation for telegraph processes (Opper and Sanguinetti, 2008):

$$\begin{aligned} KL\left[q_\nu(\nu)\|p(\nu|\chi)\right] &= \int_0^T \mathrm{d}t\, q_{\mu_i}(1,t)\left[g_{i-}\log\frac{g_{i-}}{f_{i-}} + f_{i-} - g_{i-}\right] \\ &+ q_{\mu_i}(0,t)\left[g_{i+}\log\frac{g_{i+}}{f_{i+}} + f_{i+} - g_{i+}\right]. \end{aligned} \tag{6.7}$$

The algorithm for the KL minimisation is the same as described in Chapter 5 and is based on the iteration of the following three steps: computation of the approximating diffusion process; computation of the approximating telegraph process; parameter estimation.

We start from the computation of the approximating diffusion process, which means to compute marginal mean $m(t)$ and variance $c^2(t)$ of the process. This is done by optimising the KL divergence (subject to constraints for the moments) with respect to the variational parameters $\alpha(t)$ and $\beta(t)$ using a gradient descent algorithm. By using Lagrange multipliers $\xi(t)$ and $\zeta(t)$ to incorporate the constraints for the moments (i.e. mean $m(t)$ and variance $c^2(t)$) in the KL divergence functional, we obtain the following Lagrangian (Archambeau et al., 2007)

$$\begin{aligned} \mathcal{L}\left[m, c^2, \alpha, \beta, \xi, \zeta\right] &= \langle KL\left[q_\chi(\chi)\|p(\chi|\nu)\right]\rangle_{q_\nu} - \sum_{k=1}^N \langle \log p(y_k|x(t_k))\rangle_{q_\chi} \\ &+ \int_0^T \mathrm{d}t\, \xi(t)\left[\frac{\mathrm{d}m}{\mathrm{d}t} - \alpha m - \beta\right] + \int_0^T \mathrm{d}t\, \zeta(t)\left[\frac{\mathrm{d}c^2}{\mathrm{d}t} - 2\alpha c^2 - \sigma^2\right], \end{aligned} \tag{6.8}$$

This Lagrangian can now be optimised by gradient descent with respect to $\alpha(t)$ and $\beta(t)$. We first solve forward in time the equations for the mean and the variance. Then, taking functional derivatives of the Lagrangian with respect to the moments and setting them to zero, we obtain the following ODEs for the Lagrange multipliers

$$\begin{aligned} \frac{\mathrm{d}\xi}{\mathrm{d}t} &= -\xi\alpha + \frac{1}{2\sigma^2}\,\mathrm{E}_m - \frac{1}{\sigma^2_{obs}}\sum_{k=1}^N\left[y_k - m(t)\delta(t-t_k)\right] \\ \frac{\mathrm{d}\zeta}{\mathrm{d}t} &= -2\alpha\zeta + \frac{1}{2\sigma^2}\,\mathrm{E}_{c^2} + \frac{1}{2\sigma^2_{obs}}\sum_{k=1}^N \delta(t-t_k). \end{aligned}$$

where

$$\begin{aligned} \mathrm{E}_m &= \left(\lambda^2\right)\left(4m^3(t) + 12m(t)c^2(t)\right) \\ &+ (2\lambda\alpha(t))\left(3m^2(t) + 3c^2(t)\right) \\ &+ \left(\alpha^2(t) - 2A\lambda q_\mu(1,t) - 2b\lambda + 2\beta(t)\lambda\right)(2m(t)) \\ &+ (2\alpha(t)\beta(t) - 2A\alpha(t)q_\mu(1,t) - 2b\alpha(t)) \end{aligned}$$

$$
\begin{aligned}
\mathrm{E}_{c^2} &= \left(\lambda^2\right)\left(6m^2(t)+6c^2(t)\right) \\
&+ \left(2\lambda\alpha(t)\right)\left(3m(t)\right) \\
&+ \left(\alpha^2(t)-2A\lambda q_\mu(1,t)-2b\lambda+2\beta(t)\lambda\right).
\end{aligned}
$$

These equations are solved backwards from the final condition $\xi(T)=\zeta(T)=0$. We can then compute the functional gradients of the Lagrangian with respect to the variational parameters

$$
\begin{aligned}
\frac{\delta\mathcal{L}}{\delta\alpha} &= \frac{1}{2\sigma}\left[2\lambda\left(m^3+3mc^2\right)+2\alpha\left(c^2+m^2\right)+2m\left(\beta-b\right)-2Amq_\mu(1,t)\right]-m\xi-2c^2\zeta \\
\frac{\delta\mathcal{L}}{\delta\beta} &= \frac{1}{2\sigma}\left[2\left(\beta-b\right)+2\lambda\left(m^2+c^2\right)+2m\alpha-2Aq_\mu(1,t)\right]-\xi
\end{aligned}
$$

and perform a gradient step in a gradient descent.

To compute the approximating jump process marginals, we proceed as described in Chapter 5. The marginals are only involved linearly in the KL divergence, so that fast forward-backward recursions are used for these computations (Opper et al., 2010).

The minimisation of the KL divergence represents a constrained optimisation problem where the master equation is included as a constraint through a Lagrange multiplier $\psi(t)$. The resulting Lagrangian is

$$
\begin{aligned}
\mathcal{L}[q_\mu,g_\pm,\psi] &= KL\left[q_\nu(\nu)\|p(\nu)\right]+\int \mathrm{d}t\frac{1}{2\sigma^2}[A^2-2A(\alpha m+\lambda(m^2+c^2))-2A(\beta-b)]q_\mu(1,t) \\
&+ \int \mathrm{d}t\,\psi(t)\left(\frac{\mathrm{d}q_\mu(1,t)}{\mathrm{d}t}+(g_-+g_+)q_\mu(1,t)-g_+\right), \qquad (6.9)
\end{aligned}
$$

which can be minimised as described in (Opper et al., 2010).

By inspection of Equation 6.6, we see that it is quadratic in A, b and λ, so its minimisation represents a simple quadratic programming problem which can be solved as described in Chapter 5 (Opper et al., 2010).

What we have presented in this section is essentially a theoretical extension of the work in (Opper et al., 2010) to a nonlinear system. The main advantage of this method is the fact that it provides a fast and accurate approximation so that it allows an extension to more complex gene regulatory networks as described in Chapter 5.

Of course the method presents also problems. First of all, the value of the minimum reached depends on the initial conditions. This means that for a given initial condition we obtain a unique estimation for the kinetic parameters. Multiple runs with different initialisations might provide multiple solutions. Another problem is given by the fact that we use a gradient descent algorithm, therefore a large value for the learning rate (in the gradient descent step) may speed up the convergence but also create numerical instabilities.

6.3 Introduction to the linear noise approximation

As we mentioned in Chapter 1, the chemical master equation can be approximated in different ways. Here we introduce the linear noise approximation (LNA) and we show how the promoter-mRNA stochastic system can be described in terms of a SDE using such approximation. The

LNA is derived from van Kampen's system size expansion (Van Kampen, 1981), where the CME is expanded in terms of powers of the system size Ω (i.e. the volume of the system). By neglecting terms of order $\mathcal{O}\left(\Omega^{-1}\right)$ (i.e. order of single molecule), a linear Fokker-Planck is obtained, commonly known as LNA.

We follow (Komorowski et al., 2010) and consider a stochastic system of R reactions and N species $\boldsymbol{X} = (X_1, \ldots, X_N)^T$ in a volume Ω, such that $\boldsymbol{x} = \boldsymbol{X}\Omega^{-1}$ represent the concentrations of the chemical species. We denote as $\boldsymbol{S} = \{S_{ij}\}_{i=1,\ldots,N;\, j=1,\ldots,R}$ the stoichiometric matrix and with $\hat{\boldsymbol{f}} = \{\hat{f}_j\}_{j=1,\ldots,R}$ the propensity function, such that $\hat{f}_j\,\Omega\,\Delta t$ is the transition probability for the reaction j in the infinitesimal time Δt. In particular, $\hat{f}_j(\boldsymbol{x})$ is a function of the concentrations of chemical species $\boldsymbol{x}$ and of time (which we do not represent in the notation). By introducing the step operator E, defined by

$$E^{-S_{ij}} f\left(\ldots, X_i, \ldots\right) = f\left(\ldots, X_i - S_{ij}, \ldots\right) ,$$

we can write the joint CME for the stochastic system as

$$\frac{\mathrm{d}p(\boldsymbol{X})}{\mathrm{d}t} = \Omega \sum_{j=1}^{R} \left(\prod_{i=1}^{N} E^{-S_{ij}} - 1 \right) \hat{f}_j\, p(\boldsymbol{X}) \tag{6.10}$$

where again we have simplified the notation by removing the time variable. The LNA is obtained by Taylor expanding the operator E and the propensity function $\hat{\boldsymbol{f}}$ in powers of $\Omega^{-1/2}$, around the deterministic state $\boldsymbol{\phi}$. The deterministic state $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_N)^T$ is simply obtained in the limit of an infinitely large system

$$\boldsymbol{\phi} = \lim_{\substack{\boldsymbol{X}\to\infty \\ \Omega\to\infty}} \left(\frac{\boldsymbol{X}}{\Omega} \right)$$

and is described by the so called macroscopic rate equation (MRE)

$$\frac{\mathrm{d}\phi_i}{\mathrm{d}t} = \sum_{j=1}^{R} S_{ij} f_j \qquad\qquad \text{with } \; i = 1, 2, \ldots, N\,, \tag{6.11}$$

where f_j (with $j = 1, \ldots, R$) represent the elements of the propensity function in the macroscopic limit $f_j(\boldsymbol{\phi}) = \lim_{\Omega\to\infty} \hat{f}_j(\boldsymbol{x})$. The joint probability $p(\boldsymbol{X})$ is expected to have a sharp maximum as determined by the solution of the MRE, with a width of the order $\boldsymbol{X} \sim \Omega^{1/2}$ (Van Kampen, 1981). In fact, according to Poisson statistics, when the average number of a chemical species is X_i, the fluctuations around this average is of the order $X_i^{-1/2}$ (Elf and Ehrenberg, 2003). By using this knowledge, the stochastic process X_i is decomposed as

$$X_i = \Omega\phi_i + \Omega^{1/2}\xi_i\,, \tag{6.12}$$

where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N)$ are new stochastic variables replacing $\boldsymbol{X}$. As a consequence, concentrations of the chemical species becomes $\boldsymbol{x} = \boldsymbol{\phi} + \Omega^{-1/2}\boldsymbol{\xi}$. By Taylor expanding[4] the elements of

[4]By definition, expansion of $f(x)$ at a is: $f(x) = f(a) + (1!)^{-1} f'(a)(x-a) + (2!)^{-1} f''(a)(x-a)^2 + \mathcal{O}(x^3)$.

the propensity function $\hat{f}_j(\boldsymbol{x})$ around the macroscopic value $\boldsymbol{\phi}$, we obtain

$$\begin{aligned}\hat{f}_j(\boldsymbol{x}) &= f_j(\boldsymbol{\phi}) + \sum_{i=1}^{N} \frac{\partial f_j(\boldsymbol{\phi})}{\partial \phi_i} \left(\phi_i + \Omega^{-1/2}\xi_i - \phi_i\right) + \mathcal{O}\left(\Omega^{-1}\right) \\ &= f_j(\boldsymbol{\phi}) + \Omega^{-1/2} \sum_{i=1}^{N} \frac{\partial f_j(\boldsymbol{\phi})}{\partial \phi_i} \xi_i + \mathcal{O}\left(\Omega^{-1}\right). \end{aligned} \tag{6.13}$$

We expand also the operator $\prod_{i=1}^{N} E^{-S_{ij}}$ in the CME with respect to the stochastic variable $\boldsymbol{\xi}$. As the operator $E^{-S_{ij}}$ changes X_i into $X_i - S_{ij}$, we have that it changes ξ into $\xi - \Omega^{-1/2} S_{ij}$[5]. Therefore, expansion of $\prod_{i=1}^{N} E^{-S_{ij}}$ with respect to $\boldsymbol{\xi}$ is

$$\prod_{i=1}^{N} E^{-S_{ij}} = 1 - \Omega^{-1/2} \sum_{i=1}^{N} S_{ij} \frac{\partial}{\partial \xi_i} + \frac{1}{2}\Omega^{-1} \sum_{i=1}^{N}\sum_{k=1}^{N} S_{ij} S_{kj} \frac{\partial^2}{\partial \xi_i \partial \xi_k} + \mathcal{O}\left(\Omega^{-3/2}\right). \tag{6.14}$$

The joint probability of state $\boldsymbol{\xi}$, which we denote as $\pi(\boldsymbol{\xi})$, is related to the joint $p(\boldsymbol{X})$ by the following relation

$$p\left(\boldsymbol{X}\right) = p\left(\Omega\boldsymbol{\phi} + \Omega^{1/2}\boldsymbol{\xi}\right) = \pi(\boldsymbol{\xi}).$$

Derivation of this quantity with respect to time gives the following[6]

$$\frac{\mathrm{d}p(\boldsymbol{X})}{\mathrm{d}t} = \frac{\partial \pi(\boldsymbol{\xi})}{\partial t} + \sum_{i=1}^{N} \frac{\partial \pi(\boldsymbol{\xi})}{\partial \xi_i} \frac{\mathrm{d}\xi_i}{\mathrm{d}t}. \tag{6.15}$$

By considering a constant molecules number X_i, implies

$$\frac{\mathrm{d}X_i}{\mathrm{d}t} = \Omega \frac{\mathrm{d}\phi_i}{\mathrm{d}t} + \Omega^{1/2} \frac{\mathrm{d}\xi_i}{\mathrm{d}t} = 0 \qquad \Longrightarrow \qquad \frac{\mathrm{d}\xi_i}{\mathrm{d}t} = -\Omega^{1/2} \frac{\mathrm{d}\phi_i}{\mathrm{d}t},$$

which, substituted in 6.15, gives

$$\frac{\mathrm{d}p(\boldsymbol{X})}{\mathrm{d}t} = \frac{\partial \pi(\boldsymbol{\xi})}{\partial t} - \Omega^{1/2} \sum_{i=1}^{N} \frac{\partial \pi(\boldsymbol{\xi})}{\partial \xi_i} \frac{\mathrm{d}\phi_i}{\mathrm{d}t}. \tag{6.16}$$

Now, by substituting expressions 6.16, 6.14 and 6.13 into the joint CME 6.10 we obtain the following equation

$$\begin{aligned}\frac{\partial \pi(\boldsymbol{\xi})}{\partial t} - \Omega^{1/2} \sum_{i=1}^{N} \frac{\partial \pi(\boldsymbol{\xi})}{\partial \xi_i} \frac{\mathrm{d}\phi_i}{\mathrm{d}t} &= \Omega \sum_{j=1}^{R} \left(1 - \Omega^{-1/2} \sum_{i=1}^{N} S_{ij} \frac{\partial}{\partial \xi_i} + \frac{1}{2}\Omega^{-1} \sum_{i=1}^{N}\sum_{k=1}^{N} S_{ij} S_{kj} \frac{\partial^2}{\partial \xi_i \partial \xi_k} - 1\right) \\ &\quad \left(f_j(\boldsymbol{\phi}) + \Omega^{-1/2} \sum_{i=1}^{N} \frac{\partial f_j(\boldsymbol{\phi})}{\partial \phi_i} \xi_i + \mathcal{O}\left(\Omega^{-1}\right)\right) \pi(\boldsymbol{\xi}). \end{aligned} \tag{6.17}$$

[5] In fact we have that $X_i = \Omega\phi_i + \Omega^{1/2}\xi_i$. By applying the E operator we get $X_i + 1 = \Omega\phi_i + \Omega^{1/2}\xi_i + 1 = \Omega\phi_i + \Omega^{1/2}\left(\xi_i + \Omega^{-1/2}\right)$. A change of X_i into $X_i + 1$ corresponds to a change of ξ_i into $\xi_i + \Omega^{-1/2}$ so that we have $E = 1 + \frac{\partial}{\partial X} + \frac{1}{2}\frac{\partial^2}{\partial X^2} + \mathcal{O}(3)$ and $E = 1 + \Omega^{-1/2}\frac{\partial}{\partial \xi} + \frac{1}{2}\Omega^{-1}\frac{\partial^2}{\partial \xi^2} + \mathcal{O}(3)$.

[6] Here we compute the total derivative, which takes into account of the dependency of one variable from each other. The following chain rule is used: $\frac{\mathrm{d}M}{\mathrm{d}t} = \frac{\partial M}{\partial t} + \sum_{i=1}^{N} \frac{\partial M}{\partial p_i}\frac{\mathrm{d}p_i}{\mathrm{d}t}$ where $M(p_1, \ldots, p_N)$ is a function of N variables p_i and of time as well.

Terms of order $\mathcal{O}\left(\Omega^{1/2}\right)$ on left and right hand side of 6.17 sum to zero

$$\Omega^{1/2}\sum_{i=1}^{N}\frac{\partial\pi(\boldsymbol{\xi})}{\partial\xi_i}\frac{\mathrm{d}\phi_i}{\mathrm{d}t}+\Omega^{1/2}\sum_{j=1}^{R}\sum_{i=1}^{N}S_{ij}f_j(\boldsymbol{\phi})\frac{\partial\pi(\boldsymbol{\xi})}{\partial\xi_i}=0\,,$$

as can be seen by substituting $\frac{\mathrm{d}\phi_i}{\mathrm{d}t}$ with the expression from the MRE 6.11. By disregarding of higher order terms and considering only terms of order $\mathcal{O}\left(\Omega^{0}\right)$ we obtain the following equation

$$\frac{\partial\pi(\boldsymbol{\xi})}{\partial t}=-\sum_{j=1}^{R}\left(\sum_{i,k=1}^{N}S_{ij}\frac{\partial f_j(\boldsymbol{\phi})}{\partial\phi_k}\frac{\partial\left[\xi_k\pi(\boldsymbol{\xi})\right]}{\partial\xi_i}\right)+\sum_{j=1}^{R}\left(\frac{1}{2}f_j(\boldsymbol{\phi})\sum_{i,k=1}^{N}S_{ij}S_{kj}\frac{\partial^2\pi(\boldsymbol{\xi})}{\partial\xi_i\partial\xi_k}\right)\,,\tag{6.18}$$

where we have simply changed the index of the last sum in 6.17 from i into k. We can write Equation 6.18 as

$$\frac{\partial\pi(\boldsymbol{\xi})}{\partial t}=-\sum_{i,k=1}^{N}[\boldsymbol{A}]_{ik}\frac{\partial}{\partial\xi_i}\xi_k\pi(\boldsymbol{\xi})+\frac{1}{2}\sum_{i,k=1}^{N}\left[\boldsymbol{E}\boldsymbol{E}^T\right]_{ik}\frac{\partial^2\pi(\boldsymbol{\xi})}{\partial\xi_i\partial\xi_k}\,,\tag{6.19}$$

which represents a linear Fokker-Planck equation with drift and diffusion matrix

$$[\boldsymbol{A}]_{ik} = \sum_{j=1}^{R}S_{ij}\frac{\partial f_j(\boldsymbol{\phi})}{\partial\phi_k}\tag{6.20}$$

$$\left[\boldsymbol{E}\boldsymbol{E}^T\right]_{ik} = \left[\boldsymbol{S}\,\mathrm{diag}(\boldsymbol{f}(\boldsymbol{\phi}))\,\boldsymbol{S}^T\right]_{ik}=\sum_{j=1}^{R}S_{ij}S_{kj}f_j(\boldsymbol{\phi})\,,\tag{6.21}$$

The SDE to which this Fokker-Planck equation is associated is

$$\mathrm{d}\boldsymbol{\xi}(t)=\boldsymbol{A}(t)\boldsymbol{\xi}(t)\mathrm{d}t+\boldsymbol{E}(t)\mathrm{d}\boldsymbol{w}(t)\,,\tag{6.22}$$

where $\boldsymbol{E}=\boldsymbol{S}\sqrt{\mathrm{diag}(\boldsymbol{f}(\boldsymbol{\phi}))}$. Therefore the LNA can be seen as a general method to pass from CME to SDE. Note that the variance of the diffusion is a function of $\boldsymbol{f}(\boldsymbol{\phi})$ therefore is depends on the state of the stochastic system. On the other hand, in the variational approximation described in this thesis, the variance is a constant.

The LNA can be seen as the lowest-order correction to the MRE in van Kampen's expansion. The error originated by this approximation depends on the system size Ω and can be reduced by making next order corrections. LNA can be relevant when the number of a chemical species is large, so that the fluctuations size is small compared to the species number. In this case, the discreteness of the number of species can be neglected and a continuum limit (through a SDE) becomes appropriate.

6.3.1 LNA for promoter-mRNA stochastic system

Here we describe the LNA for the promoter-mRNA linear system presented in Section 6.1. We recall that stoichiometric matrix and propensity function are given by

$$\boldsymbol{S} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \qquad \hat{\boldsymbol{f}} = \left(c_1 \frac{n_1}{\Omega},\ c_2 \frac{n_2}{\Omega},\ c_3 \frac{n_2}{\Omega},\ c_4 \frac{n_3}{\Omega} \right) .$$

By taking the limit for $\Omega \to \infty$, we obtain the propensity function in the macroscopic limit

$$\begin{aligned} \boldsymbol{f} &= (c_1\phi_1,\ c_2\phi_2,\ c_3\phi_3,\ c_4\phi_4) \\ \phi_i &= \lim_{\substack{n_i \to \infty \\ \Omega \to \infty}} \left(\frac{n_i}{\Omega} \right) \qquad \text{with } \ i = 1, 2, \ldots, N \,, \end{aligned}$$

where ϕ_i represent the deterministic states. By using these states, the stoichiometric matrix and $\boldsymbol{f}$, we obtain through Equation 6.11 the macroscopic description of the model:

$$\begin{aligned} \frac{\mathrm{d}\phi_1}{\mathrm{d}t} &= -c_1\phi_1 + c_2\phi_2 \\ \frac{\mathrm{d}\phi_2}{\mathrm{d}t} &= -c_2\phi_2 + c_1\phi_1 \\ \frac{\mathrm{d}\phi_3}{\mathrm{d}t} &= -c_4\phi_3 + c_3\phi_2 \,. \end{aligned}$$

In the LNA, deterministic and stochastic parts are decomposed according to 6.12. In order to describe the stochastic part of the model we write down drift and diffusion matrices using Equations 6.20 and 6.21:

$$\boldsymbol{A} = \begin{pmatrix} -c_1 & c_2 & 0 \\ c_1 & -c_2 & 0 \\ 0 & c_3 & -c_4 \end{pmatrix} \qquad \boldsymbol{E} = \begin{pmatrix} -\sqrt{c_1\phi_1} & \sqrt{c_2\phi_2} & 0 & 0 \\ \sqrt{c_1\phi_1} & -\sqrt{c_2\phi_2} & 0 & 0 \\ 0 & 0 & \sqrt{c_3\phi_3} & -\sqrt{c_4\phi_4} \end{pmatrix} .$$

By using these matrices into Equation 6.22, we obtain the LNA for the promoter-mRNA system in terms of SDEs:

$$\begin{aligned} \mathrm{d}\xi_1(t) &= (-c_1\xi_1 + c_2\xi_2)\mathrm{d}t + \sqrt{c_1\phi_1}\mathrm{d}w_1 + \sqrt{c_2\phi_2}\mathrm{d}w_2 \\ \mathrm{d}\xi_2(t) &= (c_1\xi_1 - c_2\xi_2)\mathrm{d}t + \sqrt{c_1\phi_1}\mathrm{d}w_1 + \sqrt{c_2\phi_2}\mathrm{d}w_2 \\ \mathrm{d}\xi_3(t) &= (c_3\xi_2 - c_4\xi_3)\mathrm{d}t + \sqrt{c_3\phi_3}\mathrm{d}w_3 + \sqrt{c_4\phi_4}\mathrm{d}w_4 \,. \end{aligned}$$

Chapter 7

Conclusions

In our understanding of complex gene regulatory mechanisms, two fundamental aspects are changing our approach to quantitative modelling. The first is the fact that we now have the techniques to measure stochasticity in gene expression. The second is that experimental data come with uncertainty, which has to be considered when using mathematical models. The combination of stochastic modelling and statistical methods represents a way to deal with these aspects.

Here, we have presented quantitative models to reverse engineer the dynamics of gene regulatory networks. Our models are based on a hybrid discrete-continuous representation in terms of Gaussian-jump processes, which enables us to describe naturally the intrinsic stochasticity in gene expression. By embedding these models in a Bayesian framework, we used them to reconstruct the profiles of latent transcription factor activities and expression of target genes. Results were obtained by adopting a variational inference approach.

This thesis showed how we have generalised a hybrid stochastic continuous-time model to construct a whole framework for studying the dynamics of gene regulatory networks. We have demonstrated that this framework has advantages over alternative approaches and can lead to useful insights into complex cellular mechanisms.

7.1 Model criticism and extension

Although our approach proved fruitful to represent gene expression data, we identify a number of aspects that can possibly limit our understanding of the gene regulation's mechanisms. These aspects are related to modelling assumptions (needed to obtain the right level of abstraction) and methodological approximations.

We have seen that a switching behaviour for TF activities is fundamental to explain allosteric modifications, which finally induce rapid actions in biological processes such as stress response (Sanguinetti et al., 2009). The use of a Markov jump process proved useful to represent the switching binary activity of transcription factors. However, as we mentioned in Chapter 4, TFs can be activated in multiple steps, due to multiple phosphorylation sites (Lee et al., 2010). In this case a 2-state Markov jump process may lack of enough flexibility to model the activity behaviour of a protein like $p53$, which is regulated in a highly complex manner.

From a similar perspective, we can conclude that modelling the TF activities as latent

variables, facilitates a description of nonlinear gene expression time-courses. However, in some cases this is not enough to explain the observed data, as we have seen in Chapter 2 for the prediction of the $PdhR$ mutants. This could be due to the fact that in our models we use fixed values for the kinetic parameters, whereas the actual values for these parameters can change according to different phases of the cell dynamics (Miller et al., 2011).

Another explanation to the unpredictability of data may be given by the presence of further components in the biological system (e.g. additional TFs, additional interactions or post-transcriptional mechanisms) which are instead neglected in our models. We also disregard of biological mechanisms such as translation and consider the whole gene regulation at the transcriptional and post-translational levels. This represents a simplification which might affect our results (especially when modelling gene regulatory networks and not purely single-layer gene structures). In fact, it has been shown that mRNA and protein concentrations are not always correlated as we assume in our models (Schwanhäusser et al., 2011; Vogel et al., 2010; Tebaldi et al., 2012).

Further simplifyng assumptions occur when modelling gene expression dynamics in eukaryotic cells. In this case, other mechanisms are involved in the regulation of gene expression. One of these is splicing, whose stochastic dynamics play a fundamental role in the production of mature mRNA molecules and has been successfully considered in biophysical models of transcription (Murugan and Kreiman, 2012). However, how to include that in our inference models, is not straightforward.

From a methodological point of view, in Chapter 5 we have used a mean field approximation for the variational distribution. Mean field approximations have been shown to be accurate for statistical inference and data assimilation problems (Eyink et al., 2004; Opper and Sanguinetti, 2008). However, when setting high levels of system noise, a decoupling of the jump processes and the Gaussian processes occurs (Opper et al., 2010). In this case, the jump process has no influence in the gene expression time-course, which is modelled by the Gaussian process alone.

Another important assumption in all of our models regards the distribution of the measurement noise, which is supposed to be Gaussian distributed. This allows a simple treatment of the inference problem, but of course it represents a limiting hypotesys on the models. The use of different noise models can be theoretically be done, by changing the jump conditions at the observations.

The development of extensions to our models is in general nontrivial. This is essentially due the fact that they have to maintain a level of abstraction such that inference is tractable. If possible extensions can be made, we would obtain a range of models which then should be compared in a statistical sense (O'Hagan, 2003).

7.2 Future perspectives

Short term plans include the development of the iterative optimisation algorithm we presented in Chapter 3 when system noise is present, as an alternative to the mean field approximation we

used in Chapter 5. We also aim to derive a variational inference approach for the case where system noise is not constant but a function of time, as in the switching model of (Stimberg et al., 2011). In this case it is not straightforward to adopt a mean field approximation as in Chapter 5. The reason is that in our mean field approximation we have assumed (to avoid infinite Kullback-Leibler divergence terms) that the approximating Gaussian process has same variance of the original process.

A careful evaluation of all the different methods for our stochastic continuous-time switching models is being developed. These methods include the conditional approximation (with and without system noise), the mean field approximation, the exact inference and the sampling based method developed by Stimberg and colleagues (Stimberg et al., 2011).

Most of the future directions for our work have been highlighted in the conclusions of Chapter 4 and Chapter 5. Further long term developments could include the incorporation in our models of epigenetic modifications (e.g. methylation and histone modifications), which have been recognised to be fundamental for the regulation of gene expression. The advent of new techniques, such as ChiP-sequencing, enabled the study of these epigenetic mechanisms from a quantitative perspective. ChiP-sequencing (ChiP-seq) is able to determine the interaction of proteins (e.g. transcription factors and chromatin-associated proteins) with the DNA sequence. These interactions have an (epigenetic) effect on the regulation of gene expression. The inclusion of these effects in our models represents an open question.

As we also mentioned in Chapter 5, another future avenue of research would be to combine our models for learning dynamics in gene regulatory networks with structure inference. So far most of the methods available for network inference use discrete-time models and in steady state condition. Therefore these methods provide information about the network interactions, which is only partial.

Appendix A

Appendix to Chapter 2

A.1 Equations for the moments of a diffusion process

We consider a multivariate diffusion process described by the following nonlinear SDE

$$\mathrm{d}\boldsymbol{x}(t) = \boldsymbol{A}(\boldsymbol{x},t)\mathrm{d}t + \boldsymbol{B}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{w}(t)\,,$$

where $\boldsymbol{A}(\boldsymbol{x},t)$ represents a $D \times 1$ vector, $\boldsymbol{B}(\boldsymbol{x},t)$ is a $D \times D$ matrix and $\boldsymbol{w}(t)$ is a multivariate Wiener process. As we described in Subsection 2.1.6, the evolution of an arbitrary function of $\boldsymbol{x}(t)$, $\boldsymbol{f}(\boldsymbol{x})$, is given by

$$\mathrm{d}\boldsymbol{f}(\boldsymbol{x}) = \left\{ \sum_i \boldsymbol{A}_i(\boldsymbol{x},t)\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_i} + \frac{1}{2}\sum_{i,j}\left[\boldsymbol{B}(\boldsymbol{x},t)\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{x},t)\right]_{ij}\frac{\partial^2 \boldsymbol{f}(\boldsymbol{x})}{\partial x_i \partial x_j} \right\}\mathrm{d}t + \sum_{i,j}\boldsymbol{B}_{ij}(\boldsymbol{x},t)\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_i}\mathrm{d}w_j\,,$$

and its expectation is

$$\mathrm{d}\left\langle \boldsymbol{f}(\boldsymbol{x})\right\rangle = \left\langle \sum_i \boldsymbol{A}_i(\boldsymbol{x},t)\frac{\partial}{\partial x_i}\boldsymbol{f}(\boldsymbol{x})\right\rangle \mathrm{d}t + \frac{1}{2}\left\langle \sum_{i,j}\left[\boldsymbol{B}(\boldsymbol{x},t)\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{x},t)\right]_{ij}\frac{\partial^2}{\partial x_i \partial x_j}\boldsymbol{f}(\boldsymbol{x})\right\rangle \mathrm{d}t\,. \tag{A.1}$$

By using this general equation we now derive equations for the first two moments of the diffusion process and then for the covariance matrix.

A.1.1 Equation for the first moment

By setting $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{x}$ we compute the two expectation terms in Equation A.1. The first term is

$$\begin{aligned}\left\langle \sum_i \boldsymbol{A}_i \frac{\partial}{\partial x_i}\boldsymbol{x}\right\rangle &= \left\langle A_1 \begin{bmatrix} \frac{\partial x_1}{\partial x_1} \\ \vdots \\ \frac{\partial x_D}{\partial x_1}\end{bmatrix} + \cdots + A_D \begin{bmatrix} \frac{\partial x_1}{\partial x_D} \\ \vdots \\ \frac{\partial x_D}{\partial x_D}\end{bmatrix}\right\rangle \\ &= \left\langle A_1 \begin{bmatrix} 1 \\ \vdots \\ 0\end{bmatrix} + \cdots + A_D \begin{bmatrix} 0 \\ \vdots \\ 1\end{bmatrix}\right\rangle = \left\langle \begin{bmatrix} A_1 \\ \vdots \\ A_D\end{bmatrix}\right\rangle = \langle \boldsymbol{A}\rangle\,,\end{aligned}$$

and the second term is

$$\left\langle \sum_{i,j} \left[\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\right]_{ij} \frac{\partial^2}{\partial x_i \partial x_j} \boldsymbol{f}(\boldsymbol{x}) \right\rangle = \left\langle \left[\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\right]_{11} \begin{bmatrix} \frac{\partial}{\partial x_1}\left(\frac{\partial}{\partial x_1} x_1\right) \\ \vdots \\ \frac{\partial}{\partial x_1}\left(\frac{\partial}{\partial x_1} x_D\right) \end{bmatrix} + \cdots + \left[\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\right]_{DD} \begin{bmatrix} \frac{\partial}{\partial x_D}\left(\frac{\partial}{\partial x_D} x_1\right) \\ \vdots \\ \frac{\partial}{\partial x_D}\left(\frac{\partial}{\partial x_D} x_D\right) \end{bmatrix} \right\rangle = 0 \,.$$

Therefore the equation for the first moment is simply

$$\frac{\mathrm{d}}{\mathrm{d}t} \langle \boldsymbol{x} \rangle = \langle \boldsymbol{A} \rangle \,. \tag{A.2}$$

A.1.2 Equation for the second moment

To find the equation for the second moment we use $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}$. The first expectation term is given by

$$\begin{aligned} \left\langle \sum_i \boldsymbol{A}_i \frac{\partial}{\partial x_i} \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \right\rangle &= \left\langle \sum_i \boldsymbol{A}_i \left[\boldsymbol{x} \frac{\partial}{\partial x_i} \boldsymbol{x}^{\mathrm{T}} + \left(\frac{\partial}{\partial x_i} \boldsymbol{x}\right) \boldsymbol{x}^{\mathrm{T}} \right] \right\rangle \\ &= \left\langle A_1 \left[\begin{pmatrix} x_1 \\ \vdots \\ x_D \end{pmatrix} (1 \cdots 0) + \begin{pmatrix} 1 \\ \vdots \\ 0 \end{pmatrix} (x_1 \cdots x_D) \right] + \cdots + \right. \\ &+ \left. A_D \left[\begin{pmatrix} x_1 \\ \vdots \\ x_D \end{pmatrix} (0 \cdots 1) + \begin{pmatrix} 0 \\ \vdots \\ 1 \end{pmatrix} (x_1 \cdots x_D) \right] \right\rangle \,. \end{aligned}$$

By solving the products we can write

$$\begin{aligned} \left\langle \sum_i \boldsymbol{A}_i \frac{\partial}{\partial x_i} \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \right\rangle &= \left\langle \begin{pmatrix} x_1 A_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ x_D A_1 & \cdots & 0 \end{pmatrix} + \begin{pmatrix} x_1 A_1 & \cdots & x_D A_1 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix} + \cdots + \right. \\ &+ \left. \begin{pmatrix} 0 & \cdots & x_1 A_D \\ \vdots & \ddots & \vdots \\ 0 & \cdots & x_D A_D \end{pmatrix} + \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ x_1 A_D & \cdots & x_D A_D \end{pmatrix} \right\rangle \,, \end{aligned}$$

and summing across the columns we obtain

$$\begin{aligned}\left\langle \sum_i \boldsymbol{A}_i \frac{\partial}{\partial x_i} \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \right\rangle &= \left\langle \begin{pmatrix} x_1 A_1 & \cdots & x_1 A_D \\ \vdots & \ddots & \vdots \\ x_D A_1 & \cdots & x_D A_D \end{pmatrix} + \begin{pmatrix} x_1 A_1 & \cdots & x_D A_1 \\ \vdots & \ddots & \vdots \\ x_1 A_D & \cdots & x_D A_D \end{pmatrix} \right\rangle \\ &= \left\langle \boldsymbol{x}\boldsymbol{A}^{\mathrm{T}} + \left(\boldsymbol{x}\boldsymbol{A}^{\mathrm{T}}\right)^{\mathrm{T}} \right\rangle = \left\langle \boldsymbol{x}\boldsymbol{A}^{\mathrm{T}} + \boldsymbol{A}\boldsymbol{x}^{\mathrm{T}} \right\rangle .\end{aligned}$$

The second term is given by

$$\begin{aligned}\left\langle \sum_{i,j} \left[\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\right]_{ij} \frac{\partial^2}{\partial x_i \partial x_j} \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \right\rangle = \Bigg\langle & \left[\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\right]_{11} \begin{pmatrix} \frac{\partial}{\partial x_1}\left(\frac{\partial}{\partial x_1} x_1^2\right) & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} \\ & + \left(\left[\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\right]_{12} + \left[\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\right]_{21}\right) \begin{pmatrix} 0 & \frac{\partial}{\partial x_1}\left(\frac{\partial}{\partial x_2} x_1 x_2\right) & \cdots & 0 \\ \frac{\partial}{\partial x_1}\left(\frac{\partial}{\partial x_2} x_1 x_2\right) & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} + \cdots + \\ & + \left[\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\right]_{DD} \begin{pmatrix} 0 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \\ 0 & \cdots & 0 & \frac{\partial}{\partial x_D}\left(\frac{\partial}{\partial x_D} x_D^2\right) \end{pmatrix} \Bigg\rangle\end{aligned}$$

which, by considering that $\left[\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\right]_{12} = \left[\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\right]_{21}$, becomes

$$\begin{aligned}\left\langle \sum_{i,j} \left[\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\right]_{ij} \frac{\partial^2}{\partial x_i \partial x_j} \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \right\rangle = \Bigg\langle & \left[\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\right]_{11} \begin{pmatrix} 2 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} \\ & + 2\left[\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\right]_{12} \begin{pmatrix} 0 & 1 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} + \cdots + \\ & + \left[\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\right]_{DD} \begin{pmatrix} 0 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \\ 0 & \cdots & 0 & 2 \end{pmatrix} \Bigg\rangle = \left\langle 2\,\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}} \right\rangle .\end{aligned}$$

Therefore the equation for the second moment is

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right\rangle = \left\langle \boldsymbol{x}\boldsymbol{A}^{\mathrm{T}} + \boldsymbol{A}\boldsymbol{x}^{\mathrm{T}}\right\rangle + \left\langle \boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\right\rangle = \left\langle \boldsymbol{x}\boldsymbol{A}^{\mathrm{T}}\right\rangle + \left\langle \boldsymbol{A}\boldsymbol{x}^{\mathrm{T}}\right\rangle + \left\langle \boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\right\rangle . \tag{A.3}$$

A.1.3 Equation for the covariance matrix

The covariance matrix is given by the following relation

$$\boldsymbol{P} = \mathbb{E}\left[(\boldsymbol{x} - \langle \boldsymbol{x}\rangle)(\boldsymbol{x} - \langle \boldsymbol{x}\rangle)^{\mathrm{T}}\right] = \left\langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right\rangle - \langle \boldsymbol{x}\rangle \langle \boldsymbol{x}\rangle^{\mathrm{T}} .$$

Then we can derive the equation for the covariance matrix as

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{P}}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\left[\left\langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right\rangle - \langle \boldsymbol{x}\rangle \langle \boldsymbol{x}\rangle^{\mathrm{T}}\right] &= \frac{\mathrm{d}}{\mathrm{d}t}\left\langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right\rangle - \left[\langle \boldsymbol{x}\rangle \frac{\mathrm{d}\langle \boldsymbol{x}\rangle^{\mathrm{T}}}{\mathrm{d}t} + \left(\frac{\mathrm{d}\langle \boldsymbol{x}\rangle}{\mathrm{d}t}\right)\langle \boldsymbol{x}\rangle^{\mathrm{T}}\right] \\ &= \frac{\mathrm{d}}{\mathrm{d}t}\left\langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right\rangle - \langle \boldsymbol{x}\rangle \left(\frac{\mathrm{d}\langle \boldsymbol{x}\rangle}{\mathrm{d}t}\right)^{\mathrm{T}} - \left(\frac{\mathrm{d}\langle \boldsymbol{x}\rangle}{\mathrm{d}t}\right)\langle \boldsymbol{x}\rangle^{\mathrm{T}} .\end{aligned}$$

Therefore, by using the equations for the moments derived above, we obtain

$$\frac{\mathrm{d}\boldsymbol{P}}{\mathrm{d}t} = \left\langle \boldsymbol{x}\boldsymbol{A}^{\mathrm{T}}\right\rangle + \left\langle \boldsymbol{A}\boldsymbol{x}^{\mathrm{T}}\right\rangle + \left\langle \boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\right\rangle - \langle \boldsymbol{x}\rangle \langle \boldsymbol{A}\rangle^{\mathrm{T}} - \langle \boldsymbol{A}\rangle \langle \boldsymbol{x}\rangle^{\mathrm{T}} . \tag{A.4}$$

A.1.4 Linear multivariate case

We now restrict to the case where

$$\begin{aligned}\boldsymbol{A}(\boldsymbol{x}, t) &= \boldsymbol{F}(t)\boldsymbol{x} + \boldsymbol{G}(t) , \\ \boldsymbol{B}(\boldsymbol{x}, t) &= \boldsymbol{B}(t) .\end{aligned}$$

The equation for the first moment becomes

$$\frac{\mathrm{d}\langle \boldsymbol{x}\rangle}{\mathrm{d}t} = \langle \boldsymbol{F}\boldsymbol{x} + \boldsymbol{G}\rangle = F\langle \boldsymbol{x}\rangle + G . \tag{A.5}$$

The equation for the covariance matrix becomes

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{P}}{\mathrm{d}t} &= \left\langle \boldsymbol{x}(\boldsymbol{F}\boldsymbol{x} + \boldsymbol{G})^{\mathrm{T}}\right\rangle + \left\langle (\boldsymbol{F}\boldsymbol{x} + \boldsymbol{G})\boldsymbol{x}^{\mathrm{T}}\right\rangle + \left\langle \boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\right\rangle - \langle \boldsymbol{x}\rangle \langle \boldsymbol{F}\boldsymbol{x} + \boldsymbol{G}\rangle^{\mathrm{T}} - \langle \boldsymbol{F}\boldsymbol{x} + \boldsymbol{G}\rangle \langle \boldsymbol{x}\rangle^{\mathrm{T}} \\ &= \left\langle \boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{F}^{\mathrm{T}} + \boldsymbol{G}^{\mathrm{T}})\right\rangle + \left\langle \boldsymbol{F}\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} + \boldsymbol{G}\boldsymbol{x}^{\mathrm{T}}\right\rangle + \boldsymbol{B}\boldsymbol{B}^{\mathrm{T}} - \langle \boldsymbol{x}\rangle \left(\langle \boldsymbol{x}\rangle^{\mathrm{T}}\boldsymbol{F}^{\mathrm{T}} + \boldsymbol{G}^{\mathrm{T}}\right) - (\boldsymbol{F}\langle \boldsymbol{x}\rangle + \boldsymbol{G})\langle \boldsymbol{x}\rangle^{\mathrm{T}} \\ &= \left(\left\langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right\rangle - \langle \boldsymbol{x}\rangle \langle \boldsymbol{x}\rangle^{\mathrm{T}}\right)\boldsymbol{F}^{\mathrm{T}} + \boldsymbol{F}\left(\left\langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right\rangle - \langle \boldsymbol{x}\rangle \langle \boldsymbol{x}\rangle^{\mathrm{T}}\right) + \boldsymbol{B}\boldsymbol{B}^{\mathrm{T}} ,\end{aligned}$$

which can be simply written as

$$\frac{\mathrm{d}\boldsymbol{P}}{\mathrm{d}t} = \boldsymbol{P}\boldsymbol{F}^{\mathrm{T}} + \boldsymbol{F}\boldsymbol{P} + \boldsymbol{B}\boldsymbol{B}^{\mathrm{T}} . \tag{A.6}$$

A.2 Update formulas for mean and variance at observation times

We derive the moments of the probability distribution given by the product of the following multivariate Gaussian distributions:

$$\begin{aligned}\mathcal{N}(\boldsymbol{z}|\boldsymbol{m_f}, \boldsymbol{C_f}) &\propto \exp\Big\{-\frac{1}{2}\left(\boldsymbol{z}-\boldsymbol{m_f}\right)^{\mathrm{T}} \boldsymbol{C_f}^{-1}\left(\boldsymbol{z}-\boldsymbol{m_f}\right)\Big\},\\ \mathcal{N}(\boldsymbol{z}|\boldsymbol{m_b}, \boldsymbol{C_b}) &\propto \exp\Big\{-\frac{1}{2}\left(\boldsymbol{z}-\boldsymbol{m_b}\right)^{\mathrm{T}} \boldsymbol{C_b}^{-1}\left(\boldsymbol{z}-\boldsymbol{m_b}\right)\Big\}.\end{aligned}$$

The product between these Gaussian distributions is proportional to

$$\begin{aligned}\mathcal{N}(\boldsymbol{z}|\boldsymbol{m_f}, \boldsymbol{C_f})\,\mathcal{N}(\boldsymbol{z}|\boldsymbol{m_b}, \boldsymbol{C_b}) &\propto \exp\Big\{-\frac{1}{2}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{C_f}^{-1}\boldsymbol{z}+\boldsymbol{z}^{\mathrm{T}}\boldsymbol{C_f}^{-1}\boldsymbol{m_f}-\frac{1}{2}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{C_b}^{-1}\boldsymbol{z}+\boldsymbol{z}^{\mathrm{T}}\boldsymbol{C_b}^{-1}\boldsymbol{m_b}\Big\}\\ &\propto \exp\Big\{-\frac{1}{2}\boldsymbol{z}^{\mathrm{T}}\left(\boldsymbol{C_f}^{-1}+\boldsymbol{C_b}^{-1}\right)\boldsymbol{z}+\boldsymbol{z}^{\mathrm{T}}\left(\boldsymbol{C_f}^{-1}\boldsymbol{m_f}+\boldsymbol{C_b}^{-1}\boldsymbol{m_b}\right)\Big\}\\ &\propto \exp\Big\{-\frac{1}{2}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{z}+\boldsymbol{z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{m}\Big\},\end{aligned}$$

where we have omitted the terms not depending on $\boldsymbol{z}$. In the last line we have used $\boldsymbol{\Sigma}$ and $\boldsymbol{m}$ to denote the covariance matrix and the mean of the resulting multivariate Gaussian. Therefore we obtain the following relation for the covariance matrix

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{C_f}^{-1} + \boldsymbol{C_b}^{-1} \;\longrightarrow\; \boldsymbol{\Sigma} = \left(\boldsymbol{C_f}^{-1} + \boldsymbol{C_b}^{-1}\right)^{-1}, \tag{A.7}$$

and for the mean

$$\begin{aligned}\boldsymbol{\Sigma}^{-1}\boldsymbol{m} = \boldsymbol{C_f}^{-1}\boldsymbol{m_f} + \boldsymbol{C_b}^{-1}\boldsymbol{m_b} \;\longrightarrow\quad \boldsymbol{m} &= \boldsymbol{\Sigma}\left(\boldsymbol{C_f}^{-1}\boldsymbol{m_f} + \boldsymbol{C_b}^{-1}\boldsymbol{m_b}\right) \qquad (A.8)\\ \boldsymbol{m} &= \left(\boldsymbol{C_f}^{-1} + \boldsymbol{C_b}^{-1}\right)^{-1}\left(\boldsymbol{C_f}^{-1}\boldsymbol{m_f} + \boldsymbol{C_b}^{-1}\boldsymbol{m_b}\right).\end{aligned}$$

In the univariate case they reduce to the following formulas

$$\sigma = \left(\frac{1}{c_f}+\frac{1}{c_b}\right)^{-1} = \frac{c_f\, c_b}{c_f + c_b} \tag{A.9}$$

$$m = \frac{c_f\, c_b}{c_f + c_b}\left(\frac{m_f}{c_f}+\frac{m_b}{c_b}\right) = \frac{m_f c_b + m_b c_f}{c_f + c_b}. \tag{A.10}$$

These formulas are needed when we compute the evolution of moments at observation times as described in Section 2.4. They are also used when we compute the product between forward and backward messages in the forward-backward algorithm.

Appendix B

Appendix to Chapter 3

B.1 Kullback-Leibler divergence between Gaussian-jump processes

Here we compute the KL divergence between a Gaussian-jump process and the variational density, as defined in Section 3.3. We compute the KL divergence as the limit for a small time increment $\Delta t \to 0$ of the following KL divergence between discretised paths:

$$\begin{aligned} KL[q(X,V)\|p(X,V)] &= \int \cdots \int \mathrm{d}x_0 \cdots \mathrm{d}x_K \sum_{\mu_0} \cdots \sum_{\mu_K} q(x_{0:K}, \mu_{0:K}) \\ &\quad \log \frac{q(x_0,\mu_0)\prod_{j=0}^{K-1} q(x_{j+1},\mu_{j+1}|x_j,\mu_j)}{p(x_0,\mu_0)\prod_{j=0}^{K-1} p(x_{j+1},\mu_{j+1}|x_j,\mu_j)} . \end{aligned}$$

To simplify the notation, we will use a singe integral symbol and a single sum symbol to represent multiple integrals and multiple sums, respectively. The variables of integration ($\mathrm{d}x_{0:K}$) and summation ($\mu_{0:K}$) will clearly indicate if we are in presence of multiple integrals:

$$KL[q(X,V)\|p(X,V)] = \int \mathrm{d}x_{0:K} \sum_{\mu_{0:K}} q(x_{0:K},\mu_{0:K}) \log \frac{q(x_0,\mu_0)\prod_{j=0}^{K-1} q(x_{j+1},\mu_{j+1}|x_j,\mu_j)}{p(x_0,\mu_0)\prod_{j=0}^{K-1} p(x_{j+1},\mu_{j+1}|x_j,\mu_j)} .$$

The previous equation can be written as

$$\begin{aligned} KL[q(X,V)\|p(X,V)] &= \int \mathrm{d}x_{0:K} \sum_{\mu_{0:K}} q(x_{0:K},\mu_{0:K}) \log \frac{q(x_0,\mu_0)}{p(x_0,\mu_0)} \\ &+ \int \mathrm{d}x_{0:K} \sum_{\mu_{0:K}} q(x_{0:K},\mu_{0:K}) \log \frac{\prod_{j=0}^{K-1} q(x_{j+1},\mu_{j+1}|x_j,\mu_j)}{\prod_{j=0}^{K-1} p(x_{j+1},\mu_{j+1}|x_j,\mu_j)} \\ &= \int \mathrm{d}x_0 \sum_{\mu_0} q(x_0,\mu_0) \log \frac{q(x_0,\mu_0)}{p(x_0,\mu_0)} \\ &+ \int \mathrm{d}x_{0:K} \sum_{\mu_{0:K}} q(x_{0:K},\mu_{0:K}) \log \frac{\prod_{j=0}^{K-1} q(x_{j+1},\mu_{j+1}|x_j,\mu_j)}{\prod_{j=0}^{K-1} p(x_{j+1},\mu_{j+1}|x_j,\mu_j)} \\ &= KL[q(x_0,\mu_0)\|p(x_0,\mu_0)] \\ &+ \int \mathrm{d}x_{0:K} \sum_{\mu_{0:K}} q(x_{0:K},\mu_{0:K}) \log \frac{\prod_{j=0}^{K-1} q(x_{j+1},\mu_{j+1}|x_j,\mu_j)}{\prod_{j=0}^{K-1} p(x_{j+1},\mu_{j+1}|x_j,\mu_j)} , \end{aligned}$$

where we have used the fact that

$$\int \mathrm{d}x_{1:K} \sum_{\mu_{1:K}} q(x_{1:K}, \mu_{1:K}) = 1\,.$$

We set $KL[q(x_0, \mu_0) \| p(x_0, \mu_0)]$ to zero and rewrite the KL divergence as follows

$$\begin{aligned} KL[q(X,V)\|p(X,V)] &= \int \mathrm{d}x_{0:K} \sum_{\mu_{0:K}} q(x_{0:K}, \mu_{0:K}) \sum_{j=0}^{K-1} \log \frac{q(x_{j+1}, \mu_{j+1}|x_j, \mu_j)}{p(x_{j+1}, \mu_{j+1}|x_j, \mu_j)} \\ &= \int \mathrm{d}x_{0:K} \sum_{\mu_{0:K}} q(x_{0:K}, \mu_{0:K}) \Big[\log \frac{q(x_1, \mu_1|x_0, \mu_0)}{p(x_1, \mu_1|x_0, \mu_0)} + \cdots + \\ &+ \log \frac{q(x_K, \mu_K|x_{K-1}, \mu_{K-1})}{p(x_K, \mu_K|x_{K-1}, \mu_{K-1})} \Big]\,. \end{aligned}$$

We separate the terms in the square brackets as

$$\begin{aligned} KL[q(X,V)\|p(X,V)] &= \int \mathrm{d}x_{0:K} \sum_{\mu_{0:K}} q(x_{0:K}, \mu_{0:K}) \log \frac{q(x_1, \mu_1|x_0, \mu_0)}{p(x_1, \mu_1|x_0, \mu_0)} + \cdots + \\ &+ \int \mathrm{d}x_{0:K} \sum_{\mu_{0:K}} q(x_{0:K}, \mu_{0:K}) \log \frac{q(x_K, \mu_K|x_{K-1}, \mu_{K-1})}{p(x_K, \mu_K|x_{K-1}, \mu_{K-1})}\,, \end{aligned}$$

and use again the property that some integrals and sums become one,

$$\begin{aligned} KL[q(X,V)\|p(X,V)] &= \int \mathrm{d}x_{0:1} \sum_{\mu_{0:1}} q(x_{0:1}, \mu_{0:1}) \log \frac{q(x_1, \mu_1|x_0, \mu_0)}{p(x_1, \mu_1|x_0, \mu_0)} + \cdots + \\ &+ \int \mathrm{d}x_{K-1:K} \sum_{\mu_{K-1:K}} q(x_{K-1:K}, \mu_{K-1:K}) \log \frac{q(x_K, \mu_K|x_{K-1}, \mu_{K-1})}{p(x_K, \mu_K|x_{K-1}, \mu_{K-1})}\,. \end{aligned}$$

The last sum of terms is equivalent to

$$\begin{aligned} KL[q(X,V)\|p(X,V)] &= \sum_{k=0}^{K-1} \int \mathrm{d}x_{k:k+1} \sum_{\mu_{k:k+1}} q(x_{k:k+1}, \mu_{k:k+1}) \log \frac{q(x_{k+1}, \mu_{k+1}|x_k, \mu_k)}{p(x_{k+1}, \mu_{k+1}|x_k, \mu_k)} \\ &= \sum_{k=0}^{K-1} \int \mathrm{d}x_{k:k+1} \sum_{\mu_{k:k+1}} q(x_{k+1}, \mu_{k+1}|x_k, \mu_k) q(x_k, \mu_k) \\ &\quad \log \frac{q(x_{k+1}, \mu_{k+1}|x_k, \mu_k)}{p(x_{k+1}, \mu_{k+1}|x_k, \mu_k)} \\ &= \sum_{k=0}^{K-1} \int \mathrm{d}x_k \sum_{\mu_k} q(x_k, \mu_k) \int \mathrm{d}x_{k+1} \sum_{\mu_{k+1}} q(x_{k+1}, \mu_{k+1}|x_k, \mu_k) \\ &\quad \log \frac{q(x_{k+1}, \mu_{k+1}|x_k, \mu_k)}{p(x_{k+1}, \mu_{k+1}|x_k, \mu_k)}\,, \end{aligned}$$

which can be simply written as

$$KL[q(X,V)\|p(X,V)] = \sum_{k=0}^{K-1} \int \mathrm{d}x_k \sum_{\mu_k} q(x_k, \mu_k) KL[q(x_{k+1}, \mu_{k+1}|x_k, \mu_k) \| p(x_{k+1}, \mu_{k+1}|x_k, \mu_k)]\,.$$

Now we use the infinitesimal transition densities as defined in Section 3.2 and 3.3:

$$p(x_{k+1},\mu_{k+1}|x_k,\mu_k) \simeq \frac{1}{\sqrt{2\pi\,\sigma^2\Delta t}}\exp\left[-\frac{(x_{k+1}-x_k-f(x_k,\mu_k)\Delta t)^2}{2\sigma^2\Delta t}\right]\left(\delta_{\mu_{k+1}\mu_k}+f(\mu_{k+1}|\mu_k,x_k)\Delta t\right)$$
$$q(x_{k+1},\mu_{k+1}|x_k,\mu_k) \simeq \frac{1}{\sqrt{2\pi\,\sigma^2\Delta t}}\exp\left[-\frac{(x_{k+1}-x_k-g(x_k,\mu_k)\Delta t)^2}{2\sigma^2\Delta t}\right]\left(\delta_{\mu_{k+1}\mu_k}+g(\mu_{k+1}|\mu_k,x_k)\Delta t\right).$$

By defining $p_{\mathcal{N}_{k+1}}$ and $q_{\mathcal{N}_{k+1}}$ to refer to the terms relative to the Gaussian parts and $p_{\mathcal{M}_{k+1}}$ and $q_{\mathcal{M}_{k+1}}$ to refer to the terms relative to the Markov jump process parts

$$\begin{aligned}
p_{\mathcal{N}_{k+1}} &= \frac{1}{\sqrt{2\pi\,\sigma^2\Delta t}}\exp\left[-\frac{(x_{k+1}-x_k-f(x_k,\mu_k)\Delta t)^2}{2\sigma^2\Delta t}\right]\\
q_{\mathcal{N}_{k+1}} &= \frac{1}{\sqrt{2\pi\,\sigma^2\Delta t}}\exp\left[-\frac{(x_{k+1}-x_k-g(x_k,\mu_k)\Delta t)^2}{2\sigma^2\Delta t}\right]\\
p_{\mathcal{M}_{k+1}} &= \left(\delta_{\mu_{k+1}\mu_k}+f(\mu_{k+1}|\mu_k,x_k)\Delta t\right)\\
q_{\mathcal{M}_{k+1}} &= \left(\delta_{\mu_{k+1}\mu_k}+g(\mu_{k+1}|\mu_k,x_k)\Delta t\right),
\end{aligned}$$

we then can write[1]:

$$\begin{aligned}
p(x_{k+1},\mu_{k+1}|x_k,\mu_k) &\simeq p_{\mathcal{N}_{k+1}}\cdot p_{\mathcal{M}_{k+1}}\\
q(x_{k+1},\mu_{k+1}|x_k,\mu_k) &\simeq q_{\mathcal{N}_{k+1}}\cdot q_{\mathcal{M}_{k+1}}.
\end{aligned}$$

Note that we are considering a general case in which both prior and posterior switching rates depend on x and, as a consequence, on the time variable. The KL divergence then can be written as

$$KL[q(X,V)\|p(X,V)] = \sum_{k=0}^{K-1}\int \mathrm{d}x_k \sum_{\mu_k} q(x_k,\mu_k)\int \mathrm{d}x_{k+1}\sum_{\mu_{k+1}} q_{\mathcal{N}_{k+1}}q_{\mathcal{M}_{k+1}}\log\frac{q_{\mathcal{N}_{k+1}}q_{\mathcal{M}_{k+1}}}{p_{\mathcal{N}_{k+1}}p_{\mathcal{M}_{k+1}}}.$$

By using the properties of the logarithm we can write

$$\begin{aligned}
KL[q(X,V)\|p(X,V)] &= \sum_{k=0}^{K-1}\int \mathrm{d}x_k \sum_{\mu_k} q(x_k,\mu_k)\left[\int \mathrm{d}x_{k+1}\sum_{\mu_{k+1}} q_{\mathcal{N}_{k+1}}q_{\mathcal{M}_{k+1}}\log\frac{q_{\mathcal{N}_{k+1}}}{p_{\mathcal{N}_{k+1}}}\right.\\
&+ \left.\int \mathrm{d}x_{k+1}\sum_{\mu_{k+1}} q_{\mathcal{N}_{k+1}}q_{\mathcal{M}_{k+1}}\log\frac{q_{\mathcal{M}_{k+1}}}{p_{\mathcal{M}_{k+1}}}\right].
\end{aligned}$$

We use the fact that we can sum $q_{\mathcal{M}_{k+1}}$ and integrate $q_{\mathcal{N}_{k+1}}$ in the first and second part of the sum, respectively,

$$\begin{aligned}
\sum_{\mu_{k+1}} q_{\mathcal{M}_{k+1}} &= 1,\\
\int \mathrm{d}x_{k+1} q_{\mathcal{N}_{k+1}} &= 1,
\end{aligned}$$

[1]The equations become exact in the limit $\Delta t \to 0$.

obtaining the following equation for the KL divergence

$$KL[q(X,V)\|p(X,V)] = \sum_{k=0}^{K-1} \int \mathrm{d}x_k \sum_{\mu_k} q(x_k,\mu_k) \left[\int \mathrm{d}x_{k+1} q_{\mathcal{N}_{k+1}} \log \frac{q_{\mathcal{N}_{k+1}}}{p_{\mathcal{N}_{k+1}}} + \sum_{\mu_{k+1}} q_{\mathcal{M}_{k+1}} \log \frac{q_{\mathcal{M}_{k+1}}}{p_{\mathcal{M}_{k+1}}} \right]. \tag{B.1}$$

B.1.1 Gaussian terms

Now we first focus on the first term in the square brackets of B.1. This represents the KL divergence between two Gaussian distributions, $\mathcal{N}(m_q, S_q) = q_{\mathcal{N}_{k+1}}$ and $\mathcal{N}(m_p, S_p) = p_{\mathcal{N}_{k+1}}$, which can be found for the multivariate case in Appendix A.5 of (Rasmussen and Williams, 2006),

$$KL[\mathcal{N}_q\|\mathcal{N}_p] = \frac{1}{2}\log\left|\boldsymbol{S}_p\boldsymbol{S}_q^{-1}\right| + \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{S}_p^{-1}\left[(\boldsymbol{m}_q-\boldsymbol{m}_p)(\boldsymbol{m}_q-\boldsymbol{m}_p)^{\mathrm{T}} + \boldsymbol{S}_q - \boldsymbol{S}_p\right]\right\}.$$

In our case we have $q_{\mathcal{N}_{k+1}} = \mathcal{N}(x_k + g(x_k,\mu_k)\Delta t, \sigma^2\Delta t)$ and $p_{\mathcal{N}_{k+1}} = \mathcal{N}(x_k + f(x_k,\mu_k)\Delta t, \sigma^2\Delta t)$, then we obtain

$$\int \mathrm{d}x_{k+1} q_{\mathcal{N}_{k+1}} \log \frac{q_{\mathcal{N}_{k+1}}}{p_{\mathcal{N}_{k+1}}} = \frac{1}{2\sigma^2}\left[f(x_k,\mu_k) - g(x_k,\mu_k)\right]^2 \Delta t\,. \tag{B.2}$$

B.1.2 Jump terms

Now we focus on the second term in the square brackets of B.1,

$$\sum_{\mu_{k+1}} q_{\mathcal{M}_{k+1}} \log \frac{q_{\mathcal{M}_{k+1}}}{p_{\mathcal{M}_{k+1}}} = \sum_{\mu_{k+1}\neq\mu_k} \left[q_{\mathcal{M}_{k+1}} \log \frac{q_{\mathcal{M}_{k+1}}}{p_{\mathcal{M}_{k+1}}} \right] + \left[q_{\mathcal{M}_{k+1}} \log \frac{q_{\mathcal{M}_{k+1}}}{p_{\mathcal{M}_{k+1}}} \right]_{\mu_{k+1}=\mu_k}$$

where we have splitted the sum over μ_{k+1} in a sum over $\mu_{k+1} \neq \mu_k$, plus a term where $\mu_{k+1} = \mu_k$. Using the equation for the transition density as defined in Section 3.1, we can rewrite the sum over $\mu_{k+1} \neq \mu_k$ and obtain

$$\sum_{\mu_{k+1}} q_{\mathcal{M}_{k+1}} \log \frac{q_{\mathcal{M}_{k+1}}}{p_{\mathcal{M}_{k+1}}} = \sum_{\mu_{k+1}\neq\mu_k} \left[g(\mu_{k+1}|\mu_k,x_k)\Delta t \log \frac{g(\mu_{k+1}|\mu_k,x_k)}{f(\mu_{k+1}|\mu_k,x_k)} \right] + \left[q_{\mathcal{M}_{k+1}} \log \frac{q_{\mathcal{M}_{k+1}}}{p_{\mathcal{M}_{k+1}}} \right]_{\mu_{k+1}=\mu_k}, \tag{B.3}$$

where the Kronecker delta terms $\delta_{\mu_{k+1}\mu_k}$ in the transition densities become zero when $\mu_{k+1} \neq \mu_k$. For the second part of the split sum, we use the normalisation property which can be derived from the transition density of a Markov jump process:

$$\sum_{\mu_{k+1}} p(\mu_{k+1}|\mu_k) \simeq \sum_{\mu_{k+1}} \left[\delta_{\mu_{k+1}\mu_k} + f(\mu_{k+1}|\mu_k)\Delta t\right].$$

The last equation becomes

$$\begin{aligned} 1 &\simeq \sum_{\mu_{k+1}\neq\mu_k} \left[\delta_{\mu_{k+1}\mu_k} + f(\mu_{k+1}|\mu_k)\Delta t\right] + \left[\delta_{\mu_{k+1}\mu_k} + f(\mu_{k+1}|\mu_k)\Delta t\right]_{\mu_{k+1}=\mu_k} \\ &\simeq \sum_{\mu_{k+1}\neq\mu_k} \left[f(\mu_{k+1}|\mu_k)\Delta t\right] + \left[1 + f(\mu_k|\mu_k)\Delta t\right] , \end{aligned}$$

which gives the following relation

$$f(\mu_k|\mu_k) = -\sum_{\mu_{k+1}\neq\mu_k} f(\mu_{k+1}|\mu_k)\,.$$

Using this relation, we can write the second term on the right hand side of B.3 as

$$\begin{aligned} \left[q_{\mathcal{M}_{k+1}} \log \frac{q_{\mathcal{M}_{k+1}}}{p_{\mathcal{M}_{k+1}}}\right]_{\mu_{k+1}=\mu_k} &= (1 + g(\mu_k|\mu_k, x_k)\Delta t) \log \frac{(1 + g(\mu_k|\mu_k, x_k)\Delta t)}{(1 + f(\mu_k|\mu_k, x_k)\Delta t)} \\ &= \left(1 - \sum_{\mu_{k+1}\neq\mu_k} g(\mu_{k+1}|\mu_k, x_k)\Delta t\right) \log \frac{\left(1 - \sum_{\mu_{k+1}\neq\mu_k} g(\mu_{k+1}|\mu_k, x_k)\Delta t\right)}{\left(1 - \sum_{\mu_{k+1}\neq\mu_k} f(\mu_{k+1}|\mu_k, x_k)\Delta t\right)} . \end{aligned}$$

Here we use the following Taylor expansion and neglect the second-order terms in Δt,

$$\log \frac{1-\gamma\Delta t}{1-\phi\Delta t} = \log(1-\gamma\Delta t) - \log(1-\phi\Delta t) \approx -\gamma\Delta t - \frac{(\gamma\Delta t)^2}{2} + \phi\Delta t + \frac{(\phi\Delta t)^2}{2} \approx -\gamma\Delta t + \phi\Delta t\,,$$

to obtain

$$\begin{aligned} \left[q_{\mathcal{M}_{k+1}} \log \frac{q_{\mathcal{M}_{k+1}}}{p_{\mathcal{M}_{k+1}}}\right]_{\mu_{k+1}=\mu_k} &\approx \left(1 - \sum_{\mu_{k+1}\neq\mu_k} g(\mu_{k+1}|\mu_k, x_k)\Delta t\right) \\ &\quad \left(-\sum_{\mu_{k+1}\neq\mu_k} g(\mu_{k+1}|\mu_k, x_k)\Delta t + \sum_{\mu_{k+1}\neq\mu_k} f(\mu_{k+1}|\mu_k, x_k)\Delta t\right) . \end{aligned}$$

Neglecting all the other second-order terms in Δt we get

$$\left[q_{\mathcal{M}_{k+1}} \log \frac{q_{\mathcal{M}_{k+1}}}{p_{\mathcal{M}_{k+1}}}\right]_{\mu_{k+1}=\mu_k} \approx \left(-\sum_{\mu_{k+1}\neq\mu_k} g(\mu_{k+1}|\mu_k, x_k)\Delta t + \sum_{\mu_{k+1}\neq\mu_k} f(\mu_{k+1}|\mu_k, x_k)\Delta t\right) .$$

Equation B.3 then becomes

$$\begin{aligned} \sum_{\mu_{k+1}} q_{\mathcal{M}_{k+1}} \log \frac{q_{\mathcal{M}_{k+1}}}{p_{\mathcal{M}_{k+1}}} &= \sum_{\mu_{k+1}\neq\mu_k} \left[g(\mu_{k+1}|\mu_k, x_k)\Delta t \log \frac{g(\mu_{k+1}|\mu_k, x_k)}{f(\mu_{k+1}|\mu_k, x_k)}\right] \\ &\quad + \left(-\sum_{\mu_{k+1}\neq\mu_k} g(\mu_{k+1}|\mu_k, x_k)\Delta t + \sum_{\mu_{k+1}\neq\mu_k} f(\mu_{k+1}|\mu_k, x_k)\Delta t\right) , \end{aligned}$$

which can be simply written as

$$\sum_{\mu_{k+1}} q_{\mathcal{M}_{k+1}} \log \frac{q_{\mathcal{M}_{k+1}}}{p_{\mathcal{M}_{k+1}}} = \sum_{\mu_{k+1}\neq\mu_k} \Big\{ g(\mu_{k+1}|\mu_k, x_k) \log \frac{g(\mu_{k+1}|\mu_k, x_k)}{f(\mu_{k+1}|\mu_k, x_k)} + f(\mu_{k+1}|\mu_k, x_k) - g(\mu_{k+1}|\mu_k, x_k) \Big\} \Delta t \,. \tag{B.4}$$

B.1.3 Final form of KL between Gaussian-jump processes

Using equations B.2 and B.4, we can rewrite the KL divergence B.1 as

$$\begin{aligned} KL[q(X,V)\|p(X,V)] &= \sum_{k=0}^{K-1} \int \mathrm{d}x_k \sum_{\mu_k} q(x_k,\mu_k) \Big[\frac{1}{2\sigma^2} \left[f(x_k,\mu_k) - g(x_k,\mu_k) \right]^2 \Delta t \\ &+ \sum_{\mu_{k+1}\neq\mu_k} \Big\{ g(\mu_{k+1}|\mu_k, x_k) \log \frac{g(\mu_{k+1}|\mu_k, x_k)}{f(\mu_{k+1}|\mu_k, x_k)} \\ &+ f(\mu_{k+1}|\mu_k, x_k) - g(\mu_{k+1}|\mu_k, x_k) \Big\} \Delta t \Big] \,. \end{aligned}$$

In the limit $\Delta t \to 0$ we obtain the KL divergence for continuous-time sample paths

$$\begin{aligned} KL[q(\chi,\nu)\|p(\chi,\nu)] &= \int_0^T \mathrm{d}t \int \mathrm{d}x \sum_{\mu} q(x,\mu,t) \Big[\frac{1}{2\sigma^2} \left[f(x,\mu) - g(x,\mu,t) \right]^2 \\ &+ \sum_{\mu'\neq\mu} \Big\{ g(\mu'|\mu,x,t) \log \frac{g(\mu'|\mu,x,t)}{f(\mu'|\mu,x,t)} + f(\mu'|\mu,x,t) - g(\mu'|\mu,x,t) \Big\} \Big] \,, \end{aligned}$$

where we have made explicit the dependence on time of the variational posterior $q(x,\mu,t)$.

B.2 Moments in the conditional approximation

Here we derive ODEs for the moments $M_1(t) = \mathbb{E}_q[x(t)]$, $M_2(t) = \mathbb{E}_q[x^2(t)]$ and $R(t) = \mathbb{E}_q[x(t)\mu(t)]$ in the conditional approximation, as defined in Section 3.4. This can be done by using the forward differential Chapman-Kolmogorov equation

$$\frac{\partial q(x,\mu,t)}{\partial t} + \frac{\partial}{\partial x}(\alpha x + B\mu + d) q(x,\mu,t) - \frac{\sigma^2}{2} \frac{\partial^2 q(x,\mu,t)}{\partial x^2} = g_\mu(t) q(x, 1-\mu, t) - g_{1-\mu}(t) q(x,\mu,t) \,, \tag{B.5}$$

where, for simplicity, we have omitted the dependence on time of the variational parameters.

B.2.1 Equation for the first moment

The equation for the first moment $M_1(t)$ can be obtained as follows:

$$\frac{\partial}{\partial t} \mathbb{E}_q[x] = \frac{\partial}{\partial t} \int \mathrm{d}x \sum_{\mu} x q(x,\mu,t) = \int \mathrm{d}x \sum_{\mu} x \left[\frac{\partial}{\partial t} q(x,\mu,t) \right] \,. \tag{B.6}$$

By using, in the last derivative, the forward differential Chapman-Kolmogorov equation, we obtain

$$\begin{aligned}\frac{\partial}{\partial t}\mathbb{E}_q[x] &= \int \mathrm{d}x \sum_\mu x \Big[-\frac{\partial}{\partial x}(\alpha x + B\mu + d)q(x,\mu,t) + \frac{\sigma^2}{2}\frac{\partial^2 q(x,\mu,t)}{\partial x^2} \\ &+ g_\mu(t)q(x,1-\mu,t) - g_{1-\mu}(t)q(x,\mu,t)\Big], \end{aligned} \tag{B.7}$$

which has to be solved with natural boundary conditions and no-flux boundary conditions (Gardiner, 2009) (sec. 5.2.1-f):

$$\begin{aligned} q(x,\mu,t) &= 0 \quad \text{as} \quad x \to \pm\infty\,, \\ \frac{\partial q(x,\mu,t)}{\partial x} &= 0 \quad \text{as} \quad x \to \pm\infty\,. \end{aligned}$$

We split the right hand side term of Equation B.7 in three bits and compute each of them. The first bit is

$$\begin{aligned}\int \mathrm{d}x \sum_\mu x \left[-\frac{\partial}{\partial x}(\alpha x + B\mu + d)q(x,\mu,t)\right] &= -\sum_\mu \Big[x(\alpha x + B\mu + d)q(x,\mu,t)\Big|_{-\infty}^{+\infty} \\ &- \int \mathrm{d}x(\alpha x + B\mu + d)q(x,\mu,t)\Big], \end{aligned}$$

where we have used integration by parts. With natural boundary conditions, the first piece is zero. The second piece represents the expectation of the drift term with respect to the posterior density:

$$\begin{aligned}\int \mathrm{d}x \sum_\mu x \left[-\frac{\partial}{\partial x}(\alpha x + B\mu + d)q(x,\mu,t)\right] &= -\sum_\mu \int \mathrm{d}x(\alpha x + B\mu + d)q(x,\mu,t) \\ &= \mathbb{E}_q[\alpha x + B\mu + d] = \alpha M_1(t) + Bq(1,t) + d\,, \end{aligned}$$

where we have used the fact that in the binary case

$$\mathbb{E}_q[\mu] = \int \mathrm{d}x \sum_\mu \mu\, q(x,\mu,t) = \int \mathrm{d}x \Big[1 \cdot q(x,1,t) + 0 \cdot q(x,0,t)\Big] = \int \mathrm{d}x\, q(x,1,t) = q(1,t)\,.$$

The second bit is

$$\int \mathrm{d}x \sum_\mu x \left[\frac{\sigma^2}{2}\frac{\partial^2 q(x,\mu,t)}{\partial x^2}\right] = \frac{\sigma^2}{2}\sum_\mu \left[x\frac{\partial q(x,\mu,t)}{\partial x}\Big|_{-\infty}^{+\infty} - \int \mathrm{d}x \frac{\partial q(x,\mu,t)}{\partial x}\right] = 0\,,$$

where we have used integration by parts. The first piece is zero, due to no-flux boundary conditions. The second term can be solved using again integration by parts

$$\int \mathrm{d}x \frac{\partial q(x,\mu,t)}{\partial x} = q(x,\mu,t)\Big|_{-\infty}^{+\infty} - \int \mathrm{d}x\, 0 \cdot q(x,\mu,t) = 0 \tag{B.8}$$

and is zero as well. Finally, the third bit of equation B.7 is

$$\int \mathrm{d}x \sum_{\mu} x \left[g_{\mu}(t) q(x, 1-\mu, t) - g_{1-\mu}(t) q(x, \mu, t) \right] = \int \mathrm{d}x \, x \left[g_1(t) q(x, 0, t) - g_0(t) q(x, 1, t) + g_0(t) q(x, 1, t) - g_1(t) q(x, 0, t) \right] , \tag{B.9}$$

where we have expressed the terms in the sum over μ. This bit becomes zero, since the terms in the square brackets sum to zero. The equation for the first moment B.7 then can be written as:

$$\frac{\mathrm{d}M_1}{\mathrm{d}t} = \alpha(t) M_1(t) + B(t) q(1, t) + d(t) . \tag{B.10}$$

B.2.2 Equation for the second moment

The equation for the second moment $M_2(t)$ can be obtained with a similar strategy:

$$\frac{\partial}{\partial t} \mathbb{E}_q[x^2] = \frac{\partial}{\partial t} \int \mathrm{d}x \sum_{\mu} x^2 q(x, \mu, t) = \int \mathrm{d}x \sum_{\mu} x^2 \left[\frac{\partial}{\partial t} q(x, \mu, t) \right] . \tag{B.11}$$

We use, in the last derivative, the forward differential Chapman-Kolmogorov equation and obtain

$$\frac{\partial}{\partial t} \mathbb{E}_q[x^2] = \int \mathrm{d}x \sum_{\mu} x^2 \Big[- \frac{\partial}{\partial x} (\alpha x + B\mu + d) q(x, \mu, t) + \frac{\sigma^2}{2} \frac{\partial^2 q(x, \mu, t)}{\partial x^2} + g_{\mu}(t) q(x, 1-\mu, t) - g_{1-\mu}(t) q(x, \mu, t) \Big] , \tag{B.12}$$

which is solved with the previous boundary conditions. We split again the right hand side term of Equation B.12 in three bits and compute each of them. The first bit is

$$\begin{aligned} \int \mathrm{d}x \sum_{\mu} x^2 \left[-\frac{\partial}{\partial x} (\alpha x + B\mu + d) q(x, \mu, t) \right] &= -\sum_{\mu} \Big[x^2 (\alpha x + B\mu + d) q(x, \mu, t) \Big|_{-\infty}^{+\infty} \\ &\quad - \int \mathrm{d}x \, 2x (\alpha x + B\mu + d) q(x, \mu, t) \Big] \\ &= \mathbb{E}_q[2x(\alpha x + B\mu + d)] \\ &= 2\alpha M_2(t) + 2BR(t) + dM_1(t) , \end{aligned}$$

where we have used integration by parts. The second bit is

$$\begin{aligned} \int \mathrm{d}x \sum_{\mu} x^2 \left[\frac{\sigma^2}{2} \frac{\partial^2 q(x, \mu, t)}{\partial x^2} \right] &= \frac{\sigma^2}{2} \sum_{\mu} \left[x^2 \frac{\partial q(x, \mu, t)}{\partial x} \Big|_{-\infty}^{+\infty} - \int \mathrm{d}x \, 2x \frac{\partial q(x, \mu, t)}{\partial x} \right] \\ &= \frac{\sigma^2}{2} \sum_{\mu} \left[-2x q(x, \mu, t) \Big|_{-\infty}^{+\infty} + \int \mathrm{d}x \, 2 q(x, \mu, t) \right] \\ &= \sigma^2 \sum_{\mu} \int \mathrm{d}x \, q(x, \mu, t) = \sigma^2 , \end{aligned}$$

where we have used integration by parts for two times. The third bit is zero; it can be easily show by replacing x with x^2 in Equation B.9. The equation for the second moment B.12 then

can be written as:

$$\frac{\mathrm{d}M_2}{\mathrm{d}t} = 2\alpha(t)M_2(t) + 2B(t)R(t) + 2d(t)M_1(t) + \sigma^2 \,. \tag{B.13}$$

B.2.3 Equation for the cross moment

Finally, the equation for the moment $R(t)$ can be obtained from

$$\frac{\partial}{\partial t}\mathbb{E}_q[x\mu] = \frac{\partial}{\partial t}\int \mathrm{d}x \sum_\mu x\mu\, q(x,\mu,t) = \int \mathrm{d}x \sum_\mu x\mu \left[\frac{\partial}{\partial t} q(x,\mu,t)\right] , \tag{B.14}$$

which, using the forward differential Chapman-Kolmogorov equation, becomes

$$\begin{aligned}\frac{\partial}{\partial t}\mathbb{E}_q[x\mu] &= \int \mathrm{d}x \sum_\mu x\mu \Big[-\frac{\partial}{\partial x}(\alpha x + B\mu + d)q(x,\mu,t) + \frac{\sigma^2}{2}\frac{\partial^2 q(x,\mu,t)}{\partial x^2} \\ &+ g_\mu(t)q(x,1-\mu,t) - g_{1-\mu}(t)q(x,\mu,t)\Big] ,\end{aligned} \tag{B.15}$$

which is solved with the natural boundary conditions and no-flux boundary conditions defined before. We split the right hand side term of Equation B.15 in three bits and compute each of them, as before. The first bit is

$$\begin{aligned}\int \mathrm{d}x \sum_\mu x\mu \left[-\frac{\partial}{\partial x}(\alpha x + B\mu + d)q(x,\mu,t)\right] &= -\sum_\mu \Big[x\mu\,(\alpha x + B\mu + d)q(x,\mu,t)\Big|_{-\infty}^{+\infty} \\ &- \int \mathrm{d}x\,\mu\,(\alpha x + B\mu + d)q(x,\mu,t)\Big] \\ &= \mathbb{E}_q[\mu(\alpha x + B\mu + d)] \\ &= \alpha R(t) + Bq(1,t) + dq(1,t)\,,\end{aligned}$$

where we have used integration by parts and the fact that in the binary case $\mu^2(t) = \mu(t)$. The second bit is

$$\int \mathrm{d}x \sum_\mu x\mu \left[\frac{\sigma^2}{2}\frac{\partial^2 q(x,\mu,t)}{\partial x^2}\right] = \frac{\sigma^2}{2}\sum_\mu \left[x\mu \frac{\partial q(x,\mu,t)}{\partial x}\Big|_{-\infty}^{+\infty} - \int \mathrm{d}x\, x \frac{\partial q(x,\mu,t)}{\partial x}\right] = 0$$

where we have used integration by parts and the resulting integral is the same as in B.8. The third bit is

$$\begin{aligned}\int \mathrm{d}x \sum_\mu x\mu \left[g_\mu(t)q(x,1-\mu,t) - g_{1-\mu}(t)q(x,\mu,t)\right] &= \int \mathrm{d}x \sum_\mu x\mu \left[g_\mu(t)q(x,t) - g_\mu(t)q(x,\mu,t)\right. \\ &- \left. g_{1-\mu}(t)q(x,\mu,t)\right] \\ &= \int \mathrm{d}x \sum_\mu x\mu \left[g_\mu(t)q(x,t) - g(t)q(x,\mu,t)\right] ,\end{aligned}$$

where we have used the relation $q(x, 1-\mu, t) = q(x, t) - q(x, \mu, t)$ (from normalisation) and the new variable $g(t) = g_\mu(t) + g_{1-\mu}(t)$. Note that $g(t)$ is independent of μ, then we obtain

$$\begin{aligned}\int \mathrm{d}x \sum_\mu x\mu \left[g_\mu(t)q(x, 1-\mu, t) - g_{1-\mu}(t)q(x, \mu, t)\right] &= \int \mathrm{d}x\, x\, g_1(t)q(x, t) - g(t) \int \mathrm{d}x \sum_\mu x\mu\, q(x, \mu, t) \\ &= g_1(t)M_1(t) - g(t)R(t)\,. \end{aligned} \tag{B.16}$$

The equation for the moment B.15 then can be written as:

$$\frac{\mathrm{d}R}{\mathrm{d}t} = [\alpha(t) - g(t)]R(t) + g_1(t)M_1(t) + [B(t) + d(t)]q(1, t)\,. \tag{B.17}$$

B.3 Derivation of moments for the combinatorial interactions case

Here we derive ODEs for the moments in the conditional approximation in the deterministic limit $\sigma \to 0$, for the combinatorial interaction case (Section 3.5). We do not use the Fokker-Planck equation, but derive the moments directly from the following:

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x}(t) = \boldsymbol{A\mu}(t) + \boldsymbol{C}\mu_1(t)\mu_2(t) + \boldsymbol{b} - \boldsymbol{\Lambda x}(t)\,. \tag{B.18}$$

For simplicity we consider the univariate case, with a single process $x(t)$:

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = A_1\mu_1(t) + A_2\mu_2(t) + C\mu_1\mu_2(t) + b - \lambda x\,, \tag{B.19}$$

where A_1, A_2 and C are scalars. Using Laplace transform, the solution of this equation is

$$x(t) = e^{-\lambda(t-t_0)}\left[x(t_0) + \int_{t_0}^{t} e^{\lambda(s-t_0)}\left[A_1\mu_1(s) + A_2\mu_2(s) + C\mu_1(s)\mu_2(s) + b\right]\mathrm{d}s\right]\,. \tag{B.20}$$

We use the initial time conditions, $t_0 = 0$ and $x(t_0) = 0$, that do not change the final result. The expectation of $x(t)$ with respect to the variational distribution, $M_1(t) = \mathbb{E}_q[x(t)]$, is given by

$$M_1(t) = e^{-\lambda t}\int_0^t e^{\lambda s}\left[A_1q_1(1, s) + A_2q_2(1, s) + Cq_1(1, s)q_2(1, s) + b\right]\mathrm{d}s\,, \tag{B.21}$$

where we have used $\mathbb{E}_q[\mu_1(t)] = q_1(1, t)$ and $\mathbb{E}_q[\mu_2(t)] = q_2(1, t)$. The ODE for the first moment[2] is then

$$\frac{\mathrm{d}M_1(t)}{\mathrm{d}t} = -\lambda M_1(t) + A_1q_1(1, t) + A_2q_{1,2}(t) + Cq_1(1, t)q_2(1, t) + b\,. \tag{B.22}$$

We now compute the ODE for the second moment. The derivative of the second moment is

$$\begin{aligned}\frac{\mathrm{d}\mathbb{E}_q[x^2(t)]}{\mathrm{d}t} &= 2\mathbb{E}_q\left[x(t)\frac{\mathrm{d}x(t)}{\mathrm{d}t}\right] \\ &= 2\mathbb{E}_q\left[x(t)\left(A_1\mu_1(t) + A_2\mu_2(t) + C\mu_1\mu_2(t) + b - \lambda x(t)\right)\right] \\ &= 2\mathbb{E}_q\left[\left(A_1x(t)\mu_1(t) + A_2x(t)\mu_2(t) + Cx(t)\mu_1\mu_2(t) + bx(t) - \lambda x^2(t)\right)\right]\,,\end{aligned}$$

[2] We first compute the derivative of the exponential $exp(-\lambda t)$ and then the derivative of the endpoint t of the integration interval.

from which we obtain the ODE for the second moment, $M_2(t) = \mathbb{E}_q[x^2(t)]$,

$$\frac{\mathrm{d}M_2(t)}{\mathrm{d}t} = -2\lambda M_2(t) + 2A_1 R_1(t) + 2A_2 R_2(t) + 2C R_{12}(t) + 2b M_1(t)\,, \tag{B.23}$$

where we have defined additional moments

$$\begin{aligned} R_1(t) &= \mathbb{E}_q[x(t)\mu_1(t)]\,, \\ R_2(t) &= \mathbb{E}_q[x(t)\mu_2(t)]\,, \\ R_{12}(t) &= \mathbb{E}_q[x(t)\mu_1(t)\mu_2(t)]\,. \end{aligned}$$

We now derive ODEs for these moments, starting from $R_1(t)$. This moment is given by the following equation

$$\begin{aligned} R_1(t) &= \mathbb{E}_q\left[\mu_1(t)x(t)\right] = \mathbb{E}_q\left[e^{-\lambda t}\int_0^t e^{\lambda s}\left[A_1\mu_1(s) + A_2\mu_2(s) + C\mu_1(s)\mu_2(s) + b\right]\mathrm{d}s \cdot \mu_1(t)\right] \\ &= e^{-\lambda t}\int_0^t e^{\lambda s} A_1 q_1(1,s,t)\mathrm{d}s + e^{-\lambda t}\int_0^t e^{\lambda s} A_2 q_1(1,t) q_2(1,s)\mathrm{d}s \\ &+ e^{-\lambda t}\int_0^t e^{\lambda s} C q_1(1,s,t) q_2(1,s)\mathrm{d}s + e^{-\lambda t}\int_0^t e^{\lambda s} b q_1(1,t)\mathrm{d}s\,, \end{aligned} \tag{B.24}$$

where we have used $\mathbb{E}_q[\mu_1(t)\mu_1(s)] = q_1(1,s,t)$ for the joint probability of the telegraph process to be in state 1 at both times s and t. By derivating $R_1(t)$ we obtain[3]

$$\begin{aligned} \frac{\mathrm{d}R_1(t)}{\mathrm{d}t} &= -\lambda R_1 + A_1 q_1(1,t) + A_2 q_1(1,t) q_2(1,t) + C q_1(1,t) q_2(1,t) + b q_1(1,t) \\ &+ e^{-\lambda t}\int_0^t e^{\lambda s} A_1(-g_1(t)) q_1(1,s,t)\mathrm{d}s + e^{-\lambda t}\int_0^t e^{\lambda s} A_1 g_{1+}(t) q_1(1,s)\mathrm{d}s \\ &+ e^{-\lambda t}\int_0^t e^{\lambda s} A_2(-g_1(t)) q_1(1,t) q_2(1,s)\mathrm{d}s + e^{-\lambda t}\int_0^t e^{\lambda s} A_2 g_{1+}(t) q_2(1,s)\mathrm{d}s \\ &+ e^{-\lambda t}\int_0^t e^{\lambda s} C(-g_1(t)) q_1(1,s,t) q_2(1,s)\mathrm{d}s + e^{-\lambda t}\int_0^t e^{\lambda s} C g_{1+}(t) q_1(1,s) q_2(1,s)\mathrm{d}s \\ &+ e^{-\lambda t}\int_0^t e^{\lambda s} b(-g_1(t)) q_1(1,t)\mathrm{d}s + e^{-\lambda t}\int_0^t e^{\lambda s} b g_{1+}(t)\mathrm{d}s\,, \end{aligned} \tag{B.25}$$

where we have used the master equation for $q_1(1,t)$ and $q_1(1,s,t)$ (see Appendix B.3.1). Grouping together the integrals on the left and grouping together the integrals on the right, we obtain the following ODE for the moment $R_1(t)$

$$\frac{\mathrm{d}R_1(t)}{\mathrm{d}t} = -[\lambda + g_1(t)]R_1(t) + [A_1 + b]q_1(1,t) + [A_2 + C]q_1(1,t)q_2(1,t) + g_{1+}(t)M_1(t)\,. \tag{B.26}$$

By symmetry we can obtain the ODE for the moment $R_2(t)$

$$\frac{\mathrm{d}R_2(t)}{\mathrm{d}t} = -[\lambda + g_2(t)]R_2(t) + [A_2 + b]q_2(1,t) + [A_1 + C]q_1(1,t)q_2(1,t) + g_{2+}(t)M_1(t)\,. \tag{B.27}$$

[3]The first term is obtained by the derivative of the exponential $exp(-\lambda t)$ and the other terms on the first line by the derivative of the endpoint t of the integration interval.

The ODE for the moment $R_{12}(t)$ can be computed with the same procedure. The moment $R_{12}(t)$ is given by

$$\begin{aligned} R_{12} &= \mathbb{E}_q\left[\mu_1(t)\mu_2(t)x(t)\right] = \mathbb{E}_q\left[e^{-\lambda t}\int_0^t e^{\lambda s}\left[A_1\mu_1(s)+A_2\mu_2(s)+C\mu_1(s)\mu_2(s)+b\right]\mathrm{d}s\cdot\mu_1(t)\mu_2(t)\right] \\ &= e^{-\lambda t}\int_0^t e^{\lambda s}A_1 q_1(1,s,t)q_2(1,t)\mathrm{d}s + e^{-\lambda t}\int_0^t e^{\lambda s}A_2 q_1(1,t)q_2(1,s,t)\mathrm{d}s \\ &+ e^{-\lambda t}\int_0^t e^{\lambda s}C q_1(1,s,t)q_2(1,s,t)\mathrm{d}s + e^{-\lambda t}\int_0^t e^{\lambda s}b q_1(1,t)q_2(1,t)\mathrm{d}s\,. \end{aligned} \tag{B.28}$$

By derivating the equation for $R_{12}(t)$ we obtain

$$\begin{aligned} \frac{\mathrm{d}R_{12}(t)}{\mathrm{d}t} &= -\lambda R_{12} + A_1 q_1(1,t)q_2(1,t) + A_2 q_1(1,t)q_2(1,t) + C q_1(1,t)q_2(1,t) + b q_1(1,t)q_2(1,t) \\ &+ e^{-\lambda t}\int_0^t e^{\lambda s}A_1(-g_1(t)q_1(1,s,t))q_2(1,t)\mathrm{d}s + e^{-\lambda t}\int_0^t e^{\lambda s}A_1 g_{1+}(t)q_1(1,s)q_2(1,t)\mathrm{d}s \\ &+ e^{-\lambda t}\int_0^t e^{\lambda s}A_2(-g_1(t)q_1(1,t))q_2(1,s,t)\mathrm{d}s + e^{-\lambda t}\int_0^t e^{\lambda s}A_2 g_{1+}(t)q_2(1,s,t)\mathrm{d}s \\ &+ e^{-\lambda t}\int_0^t e^{\lambda s}A_1(-g_1(t)q_1(1,s,t))q_2(1,t)\mathrm{d}s + e^{-\lambda t}\int_0^t e^{\lambda s}A_1 g_{1+}(t)q_1(1,s)q_2(1,t)\mathrm{d}s \\ &+ e^{-\lambda t}\int_0^t e^{\lambda s}A_1(-g_2(t)q_2(1,t))q_1(1,s,t)\mathrm{d}s + e^{-\lambda t}\int_0^t e^{\lambda s}A_1 g_{2+}(t)q_1(1,s,t)\mathrm{d}s \\ &+ e^{-\lambda t}\int_0^t e^{\lambda s}C(-g_1(t)q_1(1,s,t))q_2(1,s,t)\mathrm{d}s + e^{-\lambda t}\int_0^t e^{\lambda s}C g_{1+}(t)q_1(1,s)q_2(1,s,t)\mathrm{d}s \\ &+ e^{-\lambda t}\int_0^t e^{\lambda s}C(-g_2(t)q_2(1,s,t))q_1(1,s,t)\mathrm{d}s + e^{-\lambda t}\int_0^t e^{\lambda s}C g_{2+}(t)q_2(1,s)q_1(1,s,t)\mathrm{d}s \\ &+ e^{-\lambda t}\int_0^t e^{\lambda s}b(-g_2(t)q_2(1,t))q_1(1,t)\mathrm{d}s + e^{-\lambda t}\int_0^t e^{\lambda s}b g_{2+}(t)q_1(1,t)\mathrm{d}s \\ &+ e^{-\lambda t}\int_0^t e^{\lambda s}b(-g_1(t)q_1(1,t))q_2(1,t)\mathrm{d}s + e^{-\lambda t}\int_0^t e^{\lambda s}b g_{1+}(t)q_2(1,t)\mathrm{d}s\,, \end{aligned} \tag{B.29}$$

which, grouping together terms in $g_1(t)$, $g_2(t)$, $g_{1+}(t)$ and $g_{2+}(t)$, gives the ODE for $R_{12}(t)$:

$$\frac{\mathrm{d}R_{12}(t)}{\mathrm{d}t} = -[\lambda + g_1(t) + g_2(t)]R_{12}(t) + [A_1 + A_2 + C + b]q_1(1,t)q_2(1,t) + g_{1+}(t)R_2(t) + g_{2+}(t)R_1(t)\,.$$

B.3.1 Equations for the joint probability

We can write this joint probability as $q_1(1,s,t) = q_1(1,t|1,s)q_1(1,s)$ or $q_1(1,s,t) = q_1(1,s|1,t)q_1(1,t)$, when the order of times is $s < t$ or $s > t$, respectively. Then, in the case $s < t$, the joint probability obeys

$$\frac{\mathrm{d}q_1(1,s,t)}{\mathrm{d}t} = \frac{\mathrm{d}q_1(1,t|1,s)}{\mathrm{d}t}q_1(1,s) = \Big[-g_1(t)q_1(1,t|1,s) + g_{1+}\Big]q_1(1,s) = -g_1(t)q_1(1,s,t) + g_{1+}q_1(1,s)\,,$$

where we have defined $g_1(t) = g_{1+}(t) + g_{1-}(t)$ and we have used the fact the $q_1(1,t|1,s)$ is a function of t[4] which satisfies the master equation. In the alternative case $s > t$, the joint

[4]The conditional probability $q_1(1,t|1,s)$ represents the probability of the telegraph process to be in state 1 at time t, given that it was in state 1 at time s.

probability obeys

$$\begin{aligned}\frac{\mathrm{d}q_1(1,s,t)}{\mathrm{d}t} &= \frac{\mathrm{d}q_1(1,s|1,t)}{\mathrm{d}t}q_1(1,t) + \frac{\mathrm{d}q_1(1,t)}{\mathrm{d}t}q_1(1,s|1,t) \\ &= \Big[g_1(t)q_1(1,s|1,t) - g_{1+}(t)\Big]q_1(1,t) + \Big[-g_1(t)q_1(1,t) + g_{1+}(t)\Big]q_1(1,s|1,t) \\ &= g_{1+}(t)[q_1(1,s|1,t) - q_1(1,t)]\,,\end{aligned}$$

where now $q_1(1,s|1,t)$ and $q_1(1,t)$ are both functions of t and we have used the equation for $q_1(1,s|1,t)$. This equation can be obtained by solving the master equation for $q_1(1,t)$ for $s > t$ with initial condition $\mu_1(t) = 1$:

$$\frac{\mathrm{d}q_1(1,t|1,t)}{\mathrm{d}t} = -g_1(t)q_1(1,t|1,t) + g_{1+}(t) \quad \rightarrow \quad \frac{1}{g_1(t)}\log[g_1(t)q_1(1,t|1,t) - g_{1+}(t)]\Big|_t^s = -\,\mathrm{d}t\Big|_t^s\,,$$

where at final time s will be $q_1(1,t|1,t) = q_1(1,s|1,t)$ and at initial time $q_1(1,t|1,t) = 1$. So we obtain

$$q_1(1,s|1,t) = \frac{g_{1+}(t)}{g_1(t)} + \left(1 - \frac{g_{1+}(t)}{g_1(t)}\right)\exp\left[-g_1(t)(s-t)\right]\,,$$

which, differentiated with respect to t, gives

$$\begin{aligned}\frac{\mathrm{d}q_1(1,s|1,t)}{\mathrm{d}t} &= g_1(t)\left(1 - \frac{g_{1+}(t)}{g_1(t)}\right)\exp\left[-g_1(t)(s-t)\right] \\ &= g_1(t)\left[q_1(1,s|1,t) - \frac{g_{1+}(t)}{g_1(t)}\right] = g_1(t)q_1(1,s|1,t) - g_{1+}(t)\,.\end{aligned}$$

B.4 Optimisation for the combinatorial interactions case

Here we report the functional derivatives of the Lagrangian to optimise the variational free energy for the combinatorial interactions case. The functional derivatives with respect to the moments are the following:

$$\begin{aligned}\frac{\delta\mathcal{L}}{\delta M_{1i}(t)} &= -\frac{1}{\sigma^2_{i\,\mathrm{obs}}}\sum_{j=1}^{D} y_{ij}\delta(t-t_j) - \frac{\mathrm{d}\phi_i(t)}{\mathrm{d}t} + \lambda_i\phi_i(t) - 2b_i\kappa_i - g_{1+}(t)\psi_i(t) - g_{2+}(t)\gamma_i(t) \\ \frac{\delta\mathcal{L}}{\delta M_{2i}(t)} &= \frac{1}{2\sigma^2_{i\,\mathrm{obs}}}\sum_{j=1}^{D}\delta(t-t_j) - \frac{\mathrm{d}\kappa_i(t)}{\mathrm{d}t} + 2\lambda_i\kappa_i(t) \\ \frac{\delta\mathcal{L}}{\delta R_{1i}(t)} &= -2A_{i1}\kappa_i(t) + (\lambda_i + g_1(t))\psi_i(t) - \frac{\mathrm{d}\psi_i(t)}{\mathrm{d}t} - g_{2+}(t)\vartheta_i(t) \\ \frac{\delta\mathcal{L}}{\delta R_{2i}(t)} &= -2A_{i2}\kappa_i(t) + (\lambda_i + g_2(t))\gamma_i(t) - \frac{\mathrm{d}\gamma_i(t)}{\mathrm{d}t} - g_{1+}(t)\vartheta_i(t) \\ \frac{\delta\mathcal{L}}{\delta R_{12i}(t)} &= -2C_i\kappa_i(t) - \frac{\mathrm{d}\vartheta_i(t)}{\mathrm{d}t} + (\lambda_i + g_1(t) + g_2(t))\vartheta_i(t)\,.\end{aligned}$$

By setting these functional derivatives to zero, we obtain ODEs for the Lagrange multipliers to be solved backward with initial condition $\mathrm{multiplier}(T) = 0$ at final time T. The ODEs are solved in the order: $\kappa_i(t)$, $\vartheta_i(t)$, $\psi_i(t)$ and $\gamma_i(t)$ (or $\gamma_i(t)$ and then $\psi_i(t)$), $\phi_i(t)$. They are solved N times for $i = 1, \ldots, N$. Then we compute functional derivatives with respect to the posterior

marginals:

$$\begin{aligned}\frac{\delta\mathcal{L}}{\delta q_1(1,t)} &= \left[g_{1-}(t)\log\frac{g_{1-}(t)}{f_{1-}} + f_{1-} - g_{1-}(t)\right] - \left[g_{1+}(t)\log\frac{g_{1+}(t)}{f_{1+}} + f_{1+} - g_{1+}(t)\right] - \frac{\mathrm{d}\xi(t)}{\mathrm{d}t} \\ &+ g_1(t)\xi(t) - \sum_{i=1}^{N}[A_{i1}\phi_i(t) + C_i q_2(1,t)\phi_i(t) + (A_{i1}+b_i)\psi_i(t) + (A_{i2}+C_i)q_2(1,t)\psi_i(t) \\ &+ (A_{i1}+C_i)q_2(1,t)\gamma_i(t) + (A_{i1}+A_{i2}+C_i+b_i)q_2(1,t)\vartheta_i(t)] \qquad \text{(B.30)}\end{aligned}$$

$$\begin{aligned}\frac{\delta\mathcal{L}}{\delta q_2(1,t)} &= \left[g_{2-}(t)\log\frac{g_{2-}(t)}{f_{2-}} + f_{2-} - g_{2-}(t)\right] - \left[g_{2+}(t)\log\frac{g_{2+}(t)}{f_{2+}} + f_{2+} - g_{2+}(t)\right] - \frac{\mathrm{d}\zeta(t)}{\mathrm{d}t} \\ &+ g_2(t)\zeta(t) - \sum_{i=1}^{N}[A_{i2}\phi_i(t) + C_i q_1(1,t)\phi_i(t) + (A_{i2}+b_i)\gamma_i(t) + (A_{i1}+C_i)q_1(1,t)\gamma_i(t) \\ &+ (A_{i2}+C_i)q_1(1,t)\psi_i(t) + (A_{i1}+A_{i2}+C_i+b^i)q_1(1,t)\vartheta_i(t)\Big]\,. \qquad \text{(B.31)}\end{aligned}$$

By setting these to zero as well, we obtain ODEs for the multipliers $\xi(t)$ and $\zeta(t)$. Finally we can compute the gradients of interest, given by

$$\begin{aligned}\frac{\delta\mathcal{L}}{\delta g_{1+}(t)} &= q_1(0,t)\left(\log\frac{g_{1+}(t)}{f_{1+}} - \xi(t)\right) + \sum_{i=1}^{N}[(R_{1i}(t) - M_{1i}(t))\psi_i(t) + (R_{12i}(t) - R_{2i}(t))\vartheta_i(t)] \\ \frac{\delta\mathcal{L}}{\delta g_{2+}(t)} &= q_2(0,t)\left(\log\frac{g_{2+}(t)}{f_{2+}} - \zeta(t)\right) + \sum_{i=1}^{N}[(R_{2i}(t) - M_{1i}(t))\gamma_i(t) + (R_{12i}(t) - R_{1i}(t))\vartheta_i(t)] \\ \frac{\delta\mathcal{L}}{\delta g_{1-}(t)} &= q_1(1,t)\left(\log\frac{g_{1-}(t)}{f_{1-}} + \xi(t)\right) + \sum_{i=1}^{N}[R_{1i}(t)\psi_i(t) + R_{12i}(t)\vartheta_i(t)] \\ \frac{\delta\mathcal{L}}{\delta g_{2-}(t)} &= q_2(1,t)\left(\log\frac{g_{2-}(t)}{f_{2-}} + \zeta(t)\right) + \sum_{i=1}^{N}[R_{2i}(t)\gamma_i(t) + R_{12i}(t)\vartheta_i(t)]\,.\end{aligned}$$

These gradients are used to update the posterior switching rates in a gradient descent procedure, as described in Section 3.4.1. Estimation of the parameters is done in the same way, using the following gradients

$$\begin{aligned}\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}A_{i1}} &= \int_0^T \mathrm{d}t\Big[-\phi_i(t)q_1(1,t) - 2\kappa_i R_{1i}(t) - \psi_i(t)q_1(1,t) - (\gamma_i(t)+\vartheta_i(t))q_1(1,t)q_2(1,t)\Big] \\ \frac{\mathrm{d}\mathcal{L}}{\mathrm{d}A_{i2}} &= \int_0^T \mathrm{d}t\Big[-\phi_i(t)q_2(1,t) - 2\kappa_i(t)R_{2i}(t) - \gamma_i(t)q_2(1,t) - (\psi_i(t)+\vartheta_i(t))q_1(1,t)q_2(1,t)\Big] \\ \frac{\mathrm{d}\mathcal{L}}{\mathrm{d}C_i} &= \int_0^T \mathrm{d}t\Big[-(\phi_i(t)+\psi_i(t)+\gamma_i(t)+\vartheta_i(t))q_1(1,t)q_2(1,t) - 2\kappa_i(t)R_{12i}(t)\Big] \\ \frac{\mathrm{d}\mathcal{L}}{\mathrm{d}b_i} &= \int_0^T \mathrm{d}t\Big[-\phi_i(t) + 2\kappa_i(t)M_{1i}(t) - \psi_i(t)q_1(1,t) - \gamma_i(t)q_2(1,t) - \vartheta_i(t)q_1(1,t)q_2(1,t)\Big] \\ \frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\lambda_i} &= \int_0^T \mathrm{d}t\Big[M_{1i}(t)\phi_i(t) + 2\kappa_i(t)M_{2i}(t) + \psi_i(t)R_{1i}(t) + \gamma_i(t)R_{2i}(t) + \vartheta_i(t)R_{12i}(t)\Big]\,.\end{aligned}$$

B.5 Inference of *FNR* activity from reporter gene

The FNR activity inferred as described in Section 3.8 was compared with the FNR activity inferred from the expression levels of a reporter $lacZ$ gene. By fusing the FNR-dependent

promoter to the control region of $lacZ$, the $lacZ$ gene expression is directly linked to the FNR (promoter) activity. A high-resolution RT-PCR analysis to measure the $lacZ$ expression is performed for both anaerobic-aerobic and aerobic-anearobic transition as described in Section 3.8. Results of the inferred FNR activity are showed in Section 3.8. Here we report the posterior first moments for the five replicates used to infer the FNR activity, in both transitions (Fig. B.1).

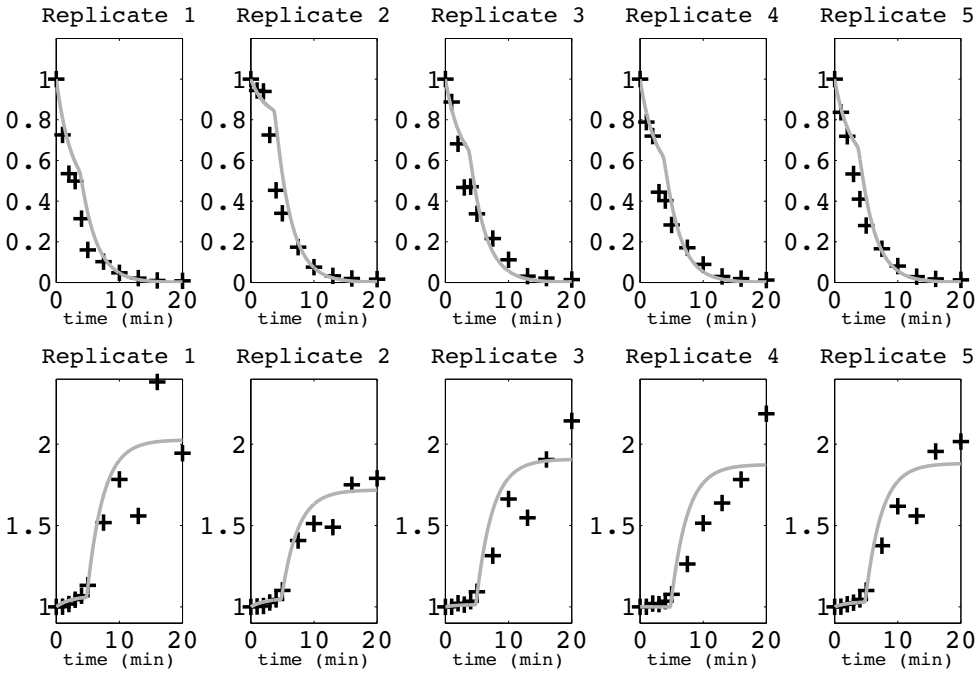

Figure B.1: Five replicates of the reporter gene expression levels. Posterior first moments obtained with approximate inference (light grey) versus noisy observations (crosses), during anaerobic-aerobic transition (upper plots) and aerobic-anaerobic transition (bottom plots). Measurements are not equivalently spaced, but taken at times t=[0,1,2,3,4,5,7.5,10,13,16,20].

Appendix C

Appendix to Chapter 4

C.1 Optimisation in FFL model (OR gate)

We report the derivation of ODEs for the moments in OR gate FFL. These ODEs make use of the approximations for the Heaviside step moments described in Subsection 4.2.2. Then we define the Lagrangian and report functional derivatives and gradients used in the optimisation procedure.

Our feed-forward loop (FFL) model is described by the following equations

$$\frac{\mathrm{d}x_s(t)}{\mathrm{d}t} = A_s\mu(t) + b_s - \lambda_s x_s(t)\,, \tag{C.1}$$

$$\frac{\mathrm{d}x_t(t)}{\mathrm{d}t} = A_t\mu(t) + b_t - \lambda_t y(t) + A_c\Theta[x_s(t) - x_c]\,, \tag{C.2}$$

where $\mu(t)$ is the latent activity of the master transcription factor, $x_s(t)$ and $x_t(t)$ represent the gene expression of the slave and target gene, respectively. The slave transcription factor is active when $x_s(t) > x_c$, where x_c is the critical threshold. Using Laplace transform, the solution of Equation (C.1) is

$$x_s(t) = e^{-\lambda_s t}\int_0^t e^{\lambda_s r}\left(A_s\mu(r) + b_s\right)\mathrm{d}r\,, \tag{C.3}$$

where we have considered $x_s(t=0) = 0$. Then, ODEs for the first and second moment can be derived as seen in Section 3.4

$$\frac{\mathrm{d}M_{1s}}{\mathrm{d}t} = -\lambda_s M_{1s} + A_s q(1,t) + b_s\,, \tag{C.4}$$

$$\frac{\mathrm{d}M_{2s}}{\mathrm{d}t} = -2\lambda_t M_{2s} + 2A_s R_s + 2b_s M_{1s}\,, \tag{C.5}$$

where

$$\frac{\mathrm{d}R_s}{\mathrm{d}t} = -(\lambda_s + g)R_s + (A_s + b_s)q(1,t) + g_+ M_{1s}\,. \tag{C.6}$$

On the other hand, the solution of Equation (C.2) is

$$x_t(t) = e^{-\lambda_t t}\left[\int_0^t e^{\lambda_t r}\left(A_t\mu(r) + b_t\right)\mathrm{d}r + \int_0^t e^{\lambda_t r}A_c\Theta[x_s(r) - x_c]\mathrm{d}r\right]\,, \tag{C.7}$$

where we have considered $x_t(t=0) = 0$ (this does not affect the following derivation). In order to calculate M_{1t} we have to compute the expectation of the Heaviside function inside

the second integral, with respect to the approximating process $q(1,t)$. By using the Laplace-type approximation[1] as defined in Subsection 4.2.2, the expectation is given by the following integral

$$\langle \Theta[x_s(t) - x_c] \rangle_q \simeq \int_{x_c}^{\infty} \mathcal{N}\left(x_s(t) | M_{1s}(t), M_{2s}(t) - M_{1s}^2(t)\right) , \tag{C.8}$$

which can be simply solved by substitution. Using the following new variable

$$m = \frac{x_s - M_{1s}}{\sqrt{2\sigma_m^2}} , \tag{C.9}$$

where $\sigma_m^2 = \left(M_{2s}(t) - M_{1s}^2(t)\right)$, the integral becomes

$$\int_{x_c}^{\infty} \mathcal{N}\left(x_s(t) | M_{1s}(t), M_{2s}(t) - M_{1s}^2(t)\right) = \int_k^{\infty} \frac{1}{\sqrt{2\pi\sigma_m^2}} e^{-m^2} \sqrt{2\sigma_m^2} \mathrm{d}m , \tag{C.10}$$

where $k = (x_c - M_{1s}(t)) \left[2\sigma_m^2\right]^{-\frac{1}{2}}$. This can be rewritten using the *error function*, as

$$\int_{x_c}^{\infty} \mathcal{N}\left(x_s(t) | M_{1s}(t), M_{2s}(t) - M_{1s}^2(t)\right) = \frac{1}{2}\left[\frac{2}{\sqrt{\pi}} \int_k^{\infty} e^{-m^2} \mathrm{d}m\right] = \frac{1}{2}\left(1 - \mathrm{erf}(k)\right) . \tag{C.11}$$

We can then obtain an equation for the first moment of $x_t(t)$

$$\frac{\mathrm{d}M_{1t}}{\mathrm{d}t} = -\lambda_t M_{1t} + (A_t q(1,t) + b_t) + \frac{1}{2} A_c \left(1 - \mathrm{erf}(k)\right) . \tag{C.12}$$

The equation for second moment of $x_t(t)$ can be obtained in this way

$$\frac{\mathrm{d}\left\langle x_t^2 \right\rangle}{\mathrm{d}t} = 2\left\langle x_t \dot{x}_t \right\rangle = 2\left\langle A_t \mu(t) x_t + b_t x_t - \lambda_t x_t^2 + A_c \Theta[x_s - x_c] \right\rangle , \tag{C.13}$$

and therefore

$$\frac{\mathrm{d}M_{2t}}{\mathrm{d}t} = -2\lambda_t M_{2t} + 2A_t R_t + 2b_t M_{1t} + 2A_c Q_t , \tag{C.14}$$

where

$$R_t = \langle \mu(t) x_t(t) \rangle , \tag{C.15}$$

$$Q_t = \langle \Theta[x_s(t) - x_c] x_t(t) \rangle , \tag{C.16}$$

represent other cross-moments. We first look at R_t which can be expressed as

$$R_t = \left\langle e^{-\lambda_t t} \left[\int_0^t e^{\lambda_t r} \left(A_t \mu(r) + b_t\right) \mathrm{d}r + \int_0^t e^{\lambda r} A_c \Theta[x_s(r) - x_c] \mathrm{d}r\right] \cdot \mu(t) \right\rangle , \tag{C.17}$$

which gives

$$R_t = e^{-\lambda_t t} \left[\int_0^t e^{\lambda_t r} \left(A_t q(1,r,t) + b_t q(1,t)\right) \mathrm{d}r + \int_0^t e^{\lambda_t r} A_c \left\langle \Theta[x_s(r) - x_c] \mu(t) \right\rangle \mathrm{d}r\right] . \tag{C.18}$$

In order to calculate R_t we have to compute the quantity $\langle \Theta[x_t(r) - x_c] \mu(t) \rangle$; assuming indepen-

[1] A short description of the general Laplace approximation is given in Appendix C.7.

dence, it can be written as $\langle\Theta[x_s(r)-x_c]\mu(t)\rangle = \langle\Theta[x_s(r)-x_c]\rangle\, q(1,t)$. In this way R_t becomes

$$R_t = e^{-\lambda_t t}\left[\int_0^t e^{\lambda_t r}\left(A_t q(1,r,t) + b_t q(1,t)\right)\mathrm{d}r + \int_0^t e^{\lambda_t r} A_c \frac{1}{2}\left(1-\mathrm{erf}(k)\right) q(1,t)\mathrm{d}r\right], \tag{C.19}$$

and deriving with respect to time we obtain

$$\frac{\mathrm{d}R_t}{\mathrm{d}t} = -(\lambda_t+g)R_t + \left(A_t + b_t + A_c\frac{1}{2}\left(1-\mathrm{erf}(k)\right)\right) q(1,t) + g_+ M_{1t}\,. \tag{C.20}$$

The quantity Q_t is given by the following expectation

$$Q_t = \left\langle e^{-\lambda_t t}\left[\int_0^t e^{\lambda_t r}\left(A_t\mu(r)+b_t\right)\mathrm{d}r + \int_0^t e^{\lambda_t r} A_c\Theta[x_s(r)-x_c]\mathrm{d}r\right]\cdot\Theta[x_s(t)-x_c]\right\rangle,$$

which gives

$$\begin{aligned} Q_t &= e^{-\lambda_t t}\left[\int_0^t e^{\lambda_t r}\left(A_t\langle\Theta[x_s(t)-x_c]\mu(r)\rangle + b_t\langle\Theta[x_s(t)-x_c]\rangle\right)\mathrm{d}r + \right. \\ &\left. \int_0^t e^{\lambda_t r} A_c\langle\Theta[x_s(r)-x_c]\Theta[x_s(t)-x_c]\rangle\,\mathrm{d}r\right]. \end{aligned} \tag{C.21}$$

This contains two more expectations to compute. The first is $\langle\Theta[x_s(t)-x_c]\mu(r)\rangle$, where the Heaviside function depends on all the history of the process, $\mu(t)$. We disregard of the dependency on the process at time r and therefore consider $\Theta[x_s(t)-x_c]$ independent of $\mu(r)$. By means of this single point approximation we can write $\langle\Theta[x_s(t)-x_c]\mu(r)\rangle = \langle\Theta[x_s(t)-x_c]\rangle\, q(r)$.

The second expectation is given by $\langle\Theta[x_s(r)-x_c]\Theta[x_s(t)-x_c]\rangle$. This quantity is equal to $\langle\Theta[x_s(r)-x_c]\rangle$ if $r=t$ (remember that the Heaviside function can be 0 or 1), and decreases to the product of the expectations $\langle\Theta[x_s(r)-x_c]\rangle\langle\Theta[x_s(t)-x_c]\rangle$ as the two Heaviside functions inside the expectation become uncorrelated. We approximate the autocorrelation function as

$$\begin{aligned} \langle\Theta[x_s(r)-x_c]\Theta[x_s(t)-x_c]\rangle &= \langle\Theta[x_s(s)-x_c]\rangle + \left(\langle\Theta[x_s(r)-x_c]\rangle\langle\Theta[x_s(t)-x_c]\rangle\right. \\ &- \left.\langle\Theta[x_s(r)-x_c]\rangle\right)\cdot\left(1-e^{-\lambda_s(t-r)}\right) \end{aligned} \tag{C.22}$$

which means that the value of the expectation decreases exponentially with the distance $t-r$ between time steps. We partially prove this, by showing that the autocorrelation of $x_s(t)$ (argument of the Heaviside function), decreases exponentially with the distance $t-r$ between time steps. The autocorrelation of $x_s(t)$ is

$$\begin{aligned} \langle x_s(t)x_s(r)\rangle &= e^{-\lambda_s(t+r)}\int_0^t\int_0^r e^{\lambda_s(\tau+\rho)}\langle\left(A_s\mu(\tau)+b_s\right)\left(A_s\mu(\rho)+b_s\right)\rangle\,\mathrm{d}\tau\mathrm{d}\rho \\ &= e^{-\lambda_s(t+r)}\int_0^t\int_0^r e^{\lambda_s(\tau+\rho)}\left\langle A_s^2\mu(\tau)\mu(\rho) + A_s b_s(\mu(\tau)+\mu(\rho)) + b_s^2\right\rangle\mathrm{d}\tau\mathrm{d}\rho\,, \end{aligned}$$

and its derivative

$$\begin{aligned} \frac{\mathrm{d}\langle x_s(t)x_s(r)\rangle}{\mathrm{d}t} &= -\lambda_s\langle x_s(t)x_s(r)\rangle \\ &+ e^{-\lambda_s(t+r)}\int_0^r e^{\lambda_s(t+\rho)}\left[A_s^2 q(1,t,\rho) + A_s b_s(q(1,t)+q(1,\rho)) + b_s^2\right]\mathrm{d}\rho\,. \end{aligned}$$

When $r = 0$, the integral in ρ (between 0 and $r = 0$) becomes null and the derivative follows the following equation

$$\frac{\mathrm{d}\left\langle x_s(t)x_s(0)\right\rangle}{\mathrm{d}t} = -\lambda_s \left\langle x_s(t)x_s(0)\right\rangle . \tag{C.23}$$

In other words, the correlation between the time 0 and the time t decays exponentially and as a consequence also the Heaviside function will decay exponentially. We showed that this is true for the times 0 and t; in order to be valid for all the times s and t, we should prove that $\langle\Theta[x_s(r) - x_c]\Theta[x_s(t) - x_c]\rangle$ is proportional to a two-dimensional error function.

Using the approximations defined above, we obtain the following ODE for Q_t

$$\begin{aligned}\frac{\mathrm{d}Q_t}{\mathrm{d}t} &= -\lambda_t Q_t + M_{1t}\frac{\mathrm{d}}{\mathrm{d}t}\left\langle\Theta[x_s(t) - x_c]\right\rangle + (A_t q(1,t) + b_t + A_c)\left\langle\Theta[x_s(t) - x_c]\right\rangle \\ &+ A_c e^{-(\lambda_s+\lambda_t)t}\left[\lambda_s\left\langle\Theta[x_s(t) - x_c]\right\rangle - \lambda_s - \frac{\mathrm{d}}{\mathrm{d}t}\left\langle\Theta[x_s(t) - x_c]\right\rangle\right] I ,\end{aligned} \tag{C.24}$$

where the integral

$$I = \int_0^t e^{(\lambda_s+\lambda_t)r}\left\langle\Theta[x_s(r) - x_c]\right\rangle \mathrm{d}r \tag{C.25}$$

is solved numerically as a Riemann integral. The derivative of the Heaviside function is computed as

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\langle\Theta[x_s(t) - x_c]\right\rangle = -\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{erf}(k) = -\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}k}\left[\frac{2}{\sqrt{\pi}}\int_0^k e^{-\tau^2}\mathrm{d}\tau\right]\frac{\mathrm{d}k}{\mathrm{d}t} , \tag{C.26}$$

which gives

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left\langle\Theta[x_s(t) - x_c]\right\rangle &= \left[\frac{x_c - M_{1s}}{2\left(M_{2s} - M_{1s}^2\right)}\left(\frac{\mathrm{d}M_{2s}}{\mathrm{d}t} - 2M_{1s}\frac{\mathrm{d}M_{1s}}{\mathrm{d}t}\right) + \frac{\mathrm{d}M_{1s}}{\mathrm{d}t}\right] \\ &\times \frac{1}{\sqrt{2\pi\left(M_{2s} - M_{1s}^2\right)}}\exp\left(-\frac{(x_c - M_{1s})^2}{2\left(M_{2s} - M_{1s}^2\right)}\right) .\end{aligned} \tag{C.27}$$

We also report functional derivatives of the expectation of Heaviside function with respect to the first and second moment of $x_s(t)$, needed to solve the backward ODEs. They are computed using as before the chain rule:

$$\begin{aligned}\frac{\delta\left\langle\Theta[x_s(t) - x_c]\right\rangle}{\delta M_{1s}} &= -\frac{1}{\sqrt{\pi}}\exp\left(-\frac{(x_c - M_{1s})^2}{2\left(M_{2s} - M_{1s}^2\right)}\right)\left(\frac{1}{\sqrt{2\left(M_{2s} - M_{1s}^2\right)}}\right)\left(\frac{M_{1s}\left(x_c - M_{1s}\right)}{M_{2s} - M_{1s}^2} - 1\right) \\ \frac{\delta\left\langle\Theta[x_s(t) - x_c]\right\rangle}{\delta M_{2s}} &= \frac{1}{\sqrt{\pi}}\exp\left(-\frac{(x_c - M_{1s})^2}{2\left(M_{2s} - M_{1s}^2\right)}\right)\frac{x_c - M_{1s}}{\left(2\left(M_{2s} - M_{1s}^2\right)\right)^{\frac{3}{2}}} .\end{aligned}$$

C.1.1 Lagrangian

Forward ODEs are incorporated into the Lagrangian, as follows

$$\begin{aligned}
\mathcal{L} &= \mathrm{KL}\left[q(\nu)||p(\nu|f_{\pm})\right] + \frac{1}{2\sigma_t^2}\sum_k \left[y_{tk}^2 - 2y_{tk}M_{1t}(t_k) + M_{2t}(t_k)\right] \\
&+ \frac{1}{2\sigma_s^2}\sum_k \left[y_{sk}^2 - 2y_{sk}M_{1s}(t_k) + M_{2s}(t_k)\right] + \int_0^T \mathrm{d}t\, \xi(t)\left[\frac{\mathrm{d}q}{\mathrm{d}t} + gq(1,t) - g_+\right] \\
&+ \int_0^T \mathrm{d}t\, \phi(t)\left[\frac{\mathrm{d}M_{1t}}{\mathrm{d}t} + \lambda_t M_{1t} - (A_t q(1,t) + b_t) - A_c\left\langle\Theta[x_s(t) - x_c]\right\rangle\right] \\
&+ \int_0^T \mathrm{d}t\, \kappa(t)\left[\frac{\mathrm{d}M_{2t}}{\mathrm{d}t} + 2\lambda_t M_{2t} - 2A_t R_t - 2b_t M_{1t} - 2A_c Q_t\right] \\
&+ \int_0^T \mathrm{d}t\, \psi(t)\left[\frac{\mathrm{d}R_t}{\mathrm{d}t} + (\lambda_t + g)R_t - \left(A_t + b_t + A_c\left\langle\Theta[x_s(t) - x_c]\right\rangle\right)q(1,t) - g_+ M_{1t}\right] \\
&+ \int_0^T \mathrm{d}t\, \gamma(t)\left[\frac{\mathrm{d}Q_t}{\mathrm{d}t} - M_{1t}\frac{\mathrm{d}}{\mathrm{d}t}\left\langle\Theta[x_s(t) - x_c]\right\rangle - (A_t q(1,t) + b_t + A_c)\left\langle\Theta[x_s(t) - x_c]\right\rangle + \right. \\
&\qquad \left. \lambda Q_t - A_c e^{-(\lambda_s+\lambda_t)t}\left[\lambda_s\left\langle\Theta[x_s(t) - x_c]\right\rangle - \lambda_s - \frac{\mathrm{d}}{\mathrm{d}t}\left\langle\Theta[x_s(t) - x_c]\right\rangle\right] \times I\right] \\
&+ \int_0^T \mathrm{d}t\, \zeta(t)\left[\frac{\mathrm{d}M_{1s}}{\mathrm{d}t} + \lambda_s M_{1s} - A_s q(1,t) - b_s\right] \\
&+ \int_0^T \mathrm{d}t\, \vartheta(t)\left[\frac{\mathrm{d}M_{2s}}{\mathrm{d}t} + 2\lambda_s M_{2s} - 2A_s R_s - 2b_s M_{1s}\right] \\
&+ \int_0^T \mathrm{d}t\, \delta(t)\left[\frac{\mathrm{d}R_s}{\mathrm{d}t} + (\lambda_s + g)R_s - (A_s + b_s)q(1,t) - g_+ M_{1s}\right]. \qquad \text{(C.28)}
\end{aligned}$$

C.1.2 Backward ODEs

Functional derivatives of the Lagrangian with respect to all moments and the variational distribution are computed to give backward ODEs for the Lagrange multipliers. We first solve the ODEs obtained by setting to zero the functional derivatives with respect to the moments of the target gene:

$$\begin{aligned}
\frac{\delta\mathcal{L}}{\delta M_{2t}} &= -\frac{\mathrm{d}\kappa}{\mathrm{d}t} + 2\lambda_t\kappa + \frac{1}{2\sigma_t^2}\sum_k \delta(t - t_k) \\
\frac{\delta\mathcal{L}}{\delta Q_t} &= -\frac{\mathrm{d}\gamma}{\mathrm{d}t} + \lambda\gamma - 2A_c\kappa \\
\frac{\delta\mathcal{L}}{\delta R_t} &= -\frac{\mathrm{d}\psi}{\mathrm{d}t} - 2A_t\kappa + (\lambda_t + g)\psi \\
\frac{\delta\mathcal{L}}{\delta M_{1t}} &= -\frac{\mathrm{d}\phi}{\mathrm{d}t} + \lambda_t\phi - 2b_t\kappa - g_+\psi - \gamma\frac{\mathrm{d}}{\mathrm{d}t}\left\langle\Theta[x_s(t) - x_c]\right\rangle - \frac{1}{\sigma_t^2}\sum_k y_{tk}\delta(t - t_k).
\end{aligned}$$

Then we compute the functional derivatives with respect to the moments of the slave gene. The functional derivatives with respect to M_{1s} and M_{2s} are a little bit tricky, due to some of the terms in the ODE for Q_t. For clarity, we only report the functional derivatives of the integral in $\mathrm{d}t$ of these four terms:

$$\gamma(t)\left[-M_{1t}\frac{\mathrm{d}\left\langle\Theta_t\right\rangle}{\mathrm{d}t} - A_c e^{-(\lambda_s+\lambda_t)t}\lambda_s\left\langle\Theta_t\right\rangle I + A_c e^{-(\lambda_s+\lambda_t)t}\lambda_s I + A_c e^{-(\lambda_s+\lambda_t)t}\frac{\mathrm{d}\left\langle\Theta_t\right\rangle}{\mathrm{d}t}I\right], \qquad \text{(C.29)}$$

where $\langle\Theta_t\rangle = \langle\Theta[x_s(t) - x_c]\rangle$. The derivative of the first term is solved integrating by parts

$$\frac{\delta}{\delta M_{1s}} \int_0^T \mathrm{d}t \; -\gamma M_{1t} \frac{\mathrm{d}\,\langle\Theta_t\rangle}{\mathrm{d}t} = \frac{\delta}{\delta M_{1s}} \left[-\langle\Theta_t\rangle \gamma M_{1t} \Big|_0^T + \int_0^T \mathrm{d}t \,\langle\Theta_t\rangle \frac{\mathrm{d}}{\mathrm{d}t}\left(\gamma M_{1t}\right) \right],$$

where the first term on the right hand side is null, due to the conditions for the Lagrange multiplier (see Subsection 3.4.1). The functional derivative then becomes[2]

$$\frac{\delta}{\delta M_{1s}} \int_0^T \mathrm{d}t \; -\gamma M_{1t} \frac{\mathrm{d}\,\langle\Theta_t\rangle}{\mathrm{d}t} = \frac{\mathrm{d}\,\langle\Theta_t\rangle}{\mathrm{d}M_{1s}} \left(M_{1t} \frac{\mathrm{d}\gamma}{\mathrm{d}t} + \gamma \frac{\mathrm{d}M_{1t}}{\mathrm{d}t} \right).$$

The functional derivative of the second term in C.29 is

$$\frac{\delta}{\delta M_{1s}} \int_0^T \mathrm{d}t \; -\gamma A_c e^{-(\lambda_s+\lambda_t)t} \lambda_s \langle\Theta_t\rangle I = -\gamma A_c e^{-(\lambda_s+\lambda_t)t} \lambda_s \left(I \frac{\delta\,\langle\Theta_t\rangle}{\delta M_{1s}} + \langle\Theta_t\rangle \frac{\delta I}{\delta M_{1s}} \right),$$

where the following integral is computed numerically

$$\frac{\delta I}{\delta M_{1s}} = \int_0^t e^{(\lambda_s+\lambda_t)r} \frac{\delta\,\langle\Theta[x_s(r) - x_c]\rangle}{\delta M_{1s}} \mathrm{d}r\,.$$

The functional derivative of the third term in C.29 is simply obtained by the previous one, considering $\langle\Theta_t\rangle = 1$ and with the opposite sign

$$\frac{\delta}{\delta M_{1s}} \int_0^T \mathrm{d}t\, \gamma A_c e^{-(\lambda_s+\lambda_t)t} \lambda_s I = \gamma A_c e^{-(\lambda_s+\lambda_t)t} \lambda_s \left(\frac{\delta I}{\delta M_{1s}} \right).$$

Finally, the functional derivative of the fourth term in C.29 is solved again integrating by parts

$$\begin{aligned} \frac{\delta}{\delta M_{1s}} \int_0^T \mathrm{d}t\, \gamma A_c e^{-(\lambda_s+\lambda_t)t} \frac{\mathrm{d}\,\langle\Theta_t\rangle}{\mathrm{d}t} I \quad &= \quad A_c \frac{\delta}{\delta M_{1s}} \left[\langle\Theta_t\rangle I \gamma e^{-(\lambda_s+\lambda_t)t} \Big|_0^T \right. \\ &- \quad \left. \int_0^T \mathrm{d}t\, \langle\Theta_t\rangle \frac{\mathrm{d}}{\mathrm{d}t} \left(I \gamma e^{-(\lambda_s+\lambda_t)t} \right) \right]. \end{aligned}$$

Again, the first term on the right hand side is null, so the remaining terms are the following

$$\begin{aligned} -A_c \frac{\delta}{\delta M_{1s}} \quad &\left[\quad \int_0^T \mathrm{d}t\, \langle\Theta_t\rangle \gamma I \left(-(\lambda_s + \lambda_t)\, e^{-(\lambda_s+\lambda_t)t} \right) \right. \\ &+ \quad \left. \int_0^T \mathrm{d}t\, \langle\Theta_t\rangle \gamma e^{-(\lambda_s+\lambda_t)t} \frac{\mathrm{d}I}{\mathrm{d}t} + \int_0^T \mathrm{d}t\, \langle\Theta_t\rangle I e^{-(\lambda_s+\lambda_t)t} \frac{\mathrm{d}\gamma}{\mathrm{d}t} \right]. \end{aligned}$$

The first bit is

$$\begin{aligned} -A_c \frac{\delta}{\delta M_{1s}} \left[\int_0^T \mathrm{d}t\, \langle\Theta_t\rangle \gamma I \left(-(\lambda_s + \lambda_t)\, e^{-(\lambda_s+\lambda_t)t} \right) \right] \quad &= \quad A_c \gamma (\lambda_s + \lambda_t) e^{-(\lambda_s+\lambda_t)t} \\ &\left(I \frac{\delta\,\langle\Theta_t\rangle}{\delta M_{1s}} + \langle\Theta_t\rangle \frac{\delta I}{\delta M_{1s}} \right). \end{aligned}$$

[2] It is the functional derivative with respect to $M_{1s}(t)$ at single time t.

Considering that the derivative of I with respect to time is $e^{(\lambda_s+\lambda_t)t}\langle\Theta_t\rangle$, the second bit becomes

$$
\begin{aligned}
-A_c\frac{\delta}{\delta M_{1s}}\left[\int_0^T \mathrm{d}t\,\langle\Theta_t\rangle\,\gamma e^{-(\lambda_s+\lambda_t)t}\frac{\mathrm{d}I}{\mathrm{d}t}\right] &= -A_c\frac{\delta}{\delta M_{1s}}\int_0^T \mathrm{d}t\,\langle\Theta_t\rangle^2\,\gamma = -A_c\gamma\frac{\delta\langle\Theta_t\rangle^2}{\delta\langle\Theta_t\rangle}\frac{\delta\langle\Theta_t\rangle}{\delta M_{1s}} \\
&= -2A_c\gamma\langle\Theta_t\rangle\frac{\delta\langle\Theta_t\rangle}{\delta M_{1s}}\,.
\end{aligned}
$$

The third bit is

$$
-A_c\frac{\delta}{\delta M_{1s}}\left[\int_0^T \mathrm{d}t\,\langle\Theta_t\rangle\,I e^{-(\lambda_s+\lambda_t)t}\frac{\mathrm{d}\gamma}{\mathrm{d}t}\right] = -A_c e^{-(\lambda_s+\lambda_t)t}\frac{\mathrm{d}\gamma}{\mathrm{d}t}\left(I\frac{\delta\langle\Theta_t\rangle}{\delta M_{1s}} + \langle\Theta_t\rangle\frac{\delta I}{\delta M_{1s}}\right).
$$

The functional derivatives of the same terms with respect to the second moment M_{2s} are computed in the same way. Then the functional derivative with respect to the second moment is

$$
\begin{aligned}
\frac{\delta\mathcal{L}}{\delta M_{2s}} &= -\frac{\mathrm{d}\vartheta}{\mathrm{d}t} + 2\lambda_s\vartheta - \frac{\mathrm{d}\langle\Theta[x_s(t)-x_c]\rangle}{\mathrm{d}M_{2s}}\left[A_c(\phi+q(1,t)\psi) + (A_t q(1,t) + b_t + A_c)\gamma\right] \\
&+ \frac{1}{2\sigma_s^2}\sum_k \delta(t-t_k) + \left[\gamma\frac{\mathrm{d}M_{1t}}{\mathrm{d}t} + M_{1t}\frac{\mathrm{d}\gamma}{\mathrm{d}t}\right]\frac{\mathrm{d}\langle\Theta[x_s(t)-x_c]\rangle}{\mathrm{d}M_{2s}} \\
&- A_c\gamma e^{-(\lambda_s+\lambda_t)t}\lambda_s\left[I\frac{\mathrm{d}\langle\Theta[x_s(t)-x_c]\rangle}{\mathrm{d}M_{2s}} + (\langle\Theta[x_s(t)-x_c]\rangle - 1)\frac{\mathrm{d}I}{\mathrm{d}M_{2s}}\right] \\
&+ A_c e^{-(\lambda_s+\lambda_t)t}\left((\lambda_s+\lambda_t)\gamma - \frac{\mathrm{d}\gamma}{\mathrm{d}t}\right)\left[I\frac{\mathrm{d}\langle\Theta[x_s(t)-x_c]\rangle}{\mathrm{d}M_{2s}} + \langle\Theta[x_s(t)-x_c]\rangle\frac{\mathrm{d}I}{\mathrm{d}M_{2s}}\right] \\
&- 2A_c\gamma\langle\Theta[x_s(t)-x_c]\rangle\frac{\mathrm{d}\langle\Theta[x_s(t)-x_c]\rangle}{\mathrm{d}M_{2s}}\,.
\end{aligned}
$$

The functional derivatives with respect to the cross moment R_s is the same we have seen in Subsection 3.4.1

$$
\frac{\delta\mathcal{L}}{\delta R_s} = -\frac{\mathrm{d}\delta}{\mathrm{d}t} + (\lambda_s+g)\delta - 2A_s\vartheta\,,
$$

and the functional derivatives with respect to the first moment is

$$
\begin{aligned}
\frac{\delta\mathcal{L}}{\delta M_{1s}} &= -\frac{\mathrm{d}\zeta}{\mathrm{d}t} + \lambda_s\zeta - g_+\delta - \frac{\mathrm{d}\langle\Theta[x_s(t)-x_c]\rangle}{\mathrm{d}M_{1s}}\left[A_c(\phi+q(1,t)\psi) + (A_t q(1,t) + b_t + A_c)\gamma\right] \\
&- \frac{1}{\sigma_s^2}\sum_k y_{sk}\delta(t-t_k) - 2b_s\vartheta + \left[\gamma\frac{\mathrm{d}M_{1t}}{\mathrm{d}t} + M_{1t}\frac{\mathrm{d}\gamma}{\mathrm{d}t}\right]\frac{\mathrm{d}\langle\Theta[x_s(t)-x_c]\rangle}{\mathrm{d}M_{1s}} \\
&- A_c\gamma e^{-(\lambda_s+\lambda_t)t}\lambda_s\left[I\frac{\mathrm{d}\langle\Theta[x_s(t)-x_c]\rangle}{\mathrm{d}M_{1s}} + (\langle\Theta[x_s(t)-x_c]\rangle - 1)\frac{\mathrm{d}I}{\mathrm{d}M_{1s}}\right] \\
&+ A_c e^{-(\lambda_s+\lambda_t)t}\left((\lambda_s+\lambda_t)\gamma - \frac{\mathrm{d}\gamma}{\mathrm{d}t}\right)\left[I\frac{\mathrm{d}\langle\Theta[x_s(t)-x_c]\rangle}{\mathrm{d}M_{1s}} + \langle\Theta[x_s(t)-x_c]\rangle\frac{\mathrm{d}I}{\mathrm{d}M_{1s}}\right] \\
&- 2A_c\gamma\langle\Theta[x_s(t)-x_c]\rangle\frac{\mathrm{d}\langle\Theta[x_s(t)-x_c]\rangle}{\mathrm{d}M_{1s}}\,.
\end{aligned}
$$

Finally, the functional derivative with respect to the variational distribution is

$$
\begin{aligned}
\frac{\delta\mathcal{L}}{\delta q} &= \left[g_-\log\frac{g_-}{f_-} + f_- - g_-\right] - \left[g_+\log\frac{g_+}{f_+} + f_+ - g_+\right] - \frac{\mathrm{d}\xi}{\mathrm{d}t} + g\xi - A_t\phi \\
&- (A_t + b_t + A_c\langle\Theta[x_s(t)-x_c]\rangle)\psi - \gamma A_t\langle\Theta[x_s(t)-x_c]\rangle - \zeta A_s - \delta(A_s+b_s)\,. \qquad \text{(C.30)}
\end{aligned}
$$

By setting to zero all the functional derivatives we obtain ODEs for the Lagrange multipliers. These are solved backward in the order we have reported them using Runge-Kutta method

(RK4), with value 0 for the multipliers at final time T.

C.1.3 Gradients

We report the gradients for the update of the switching rates of the master TF and the kinetic parameters of both slave and target gene. Gradients for the transition rates are:

$$\frac{\delta \mathcal{L}}{\delta g_+} = (1-q(1,t))\log\frac{g_+}{f_+} + \xi(q(1,t)-1) + \psi(R_t - M_{1t}) + \delta(R_s - M_{1s})\,, \tag{C.31}$$

$$\frac{\delta \mathcal{L}}{\delta g_-} = q(1,t)\log\frac{g_-}{f_-} + \xi q(1,t) + \psi R_t + \delta R_s\,. \tag{C.32}$$

The gradients with respect to all parameters are computed together with the previous gradients during the optimisation algorithm. Gradients with respect to parameters of the slave gene x_s are given by:

$$\frac{\delta \mathcal{L}}{\delta A_s} = -\int_0^T \mathrm{d}t\,[\zeta q(1,t) + 2\vartheta R_s + \delta q(1,t)]\,, \tag{C.33}$$

$$\frac{\delta \mathcal{L}}{\delta b_s} = -\int_0^T \mathrm{d}t\,[\zeta + 2\vartheta M_{1s} + \delta q(1,t)]\,, \tag{C.34}$$

$$\frac{\delta \mathcal{L}}{\delta \lambda_s} = \int_0^T \mathrm{d}t\,[\zeta M_{1s} + 2\vartheta M_{2s} + \delta R_s]\,, \tag{C.35}$$

where we have disregarded of the functional derivative of terms with respect to λ_s when they have originated from approximation (4.15). The gradient with respect to parameter A_c is

$$\begin{aligned}\frac{\delta \mathcal{L}}{\delta A_c} &= -\int_0^T \mathrm{d}t\Big[\phi\,\langle\Theta[x_s(t)-x_c]\rangle + 2\kappa Q_t + \psi\,\langle\Theta[x(t)-c]\rangle\,q(t) \\ &+ \gamma\left(\langle\Theta[x(t)-c]\rangle + e^{-(\lambda+\lambda_1)t}\left[\lambda_1\,\langle\Theta[x(t)-c]\rangle - \lambda_1 - \frac{\mathrm{d}}{\mathrm{d}t}\,\langle\Theta[x(t)-c]\rangle\right]\times I\right)\Big]\,.\end{aligned} \tag{C.36}$$

Finally, gradients with respect to the other parameters are given by

$$\frac{\delta \mathcal{L}}{\delta A_t} = -\int_0^T \mathrm{d}t\,[\phi q(1,t) + 2\kappa R_t + \psi q(1,t) + \gamma\,\langle\Theta[x_s(t)-x_c]\rangle\,q(1,t)]\,, \tag{C.37}$$

$$\frac{\delta \mathcal{L}}{\delta b_t} = -\int_0^T \mathrm{d}t\,[\phi + 2\kappa M_{1t} + \psi q(1,t) + \gamma\,\langle\Theta[x_s(t)-x_c]\rangle]\,, \tag{C.38}$$

$$\begin{aligned}\frac{\delta \mathcal{L}}{\delta \lambda_t} &= \int_0^T \mathrm{d}t\Big[\phi M_{1t} + 2\kappa M_{2t} + \psi R_t + \gamma Q_t \\ &+ \gamma A_c e^{-(\lambda_s+\lambda_t)t}\left[\lambda_s\left(\langle\Theta[x_s(t)-x_c]\rangle - 1\right) - \frac{\mathrm{d}\,\langle\Theta[x_s(t)-x_c]\rangle}{\mathrm{d}t}\right] \\ &\times \left\{t\int_0^t e^{(\lambda_s+\lambda_t)r}\,\langle\Theta[x_s(r)-x_c]\rangle\,\mathrm{d}r - \int_0^t r\,e^{(\lambda_s+\lambda_t)r}\,\langle\Theta[x_s(r)-x_c]\rangle\,\mathrm{d}r\right\}\Big]\,.\end{aligned} \tag{C.39}$$

C.2 Optimisation in FFL model (AND gate)

Here we report ODEs for the moments and cross-moments of the target gene in our AND gate FFL model. The model is described by the following equations

$$\frac{\mathrm{d}x_s(t)}{\mathrm{d}t} = A_s\mu(t) + b_s - \lambda_s x_s(t)\,, \tag{C.40}$$

$$\frac{\mathrm{d}x_t(t)}{\mathrm{d}t} = A_t\mu(t)\Theta[x_s(t) - x_c] + b_t - \lambda_t x_t(t)\,. \tag{C.41}$$

Using Laplace transform, the solution of Equation C.41 is

$$x_t(t) = e^{-\lambda_t t}\left[\int_0^t e^{\lambda_t r}\left(A_t\mu(r)\Theta[x_s(r) - x_c] + b_t\right)\mathrm{d}r\right], \tag{C.42}$$

and so, in order to calculate $M_{1t} = \langle x_t\rangle$ we have to compute the expectation of the quantity $\mu(r)\Theta[x_s(r)-x_c]$. We use the approximation described above and assume independence between the two terms (this implies factorisation). The independence is justified by the fact that μ changes rapidly compared to Θ; it becomes exact when the frequency of the switchings of μ is very high. Then we obtain an equation for the first moment of $x_t(t)$

$$\frac{\mathrm{d}M_{1t}}{\mathrm{d}t} = -\lambda_t M_{1t} + A_t q(1,t)\frac{1}{2}\left(1 - \mathrm{erf}(k)\right) + b_t\,, \tag{C.43}$$

where k is the same as described above. The equation for second moment of $x_t(t)$ can be obtained from

$$\frac{\mathrm{d}\left\langle x_t^2\right\rangle}{\mathrm{d}t} = 2\left\langle x_t\dot{x}_t\right\rangle = 2\left\langle A_t\mu(t)\Theta[x_s(t) - x_c]x_t + b_t x_t - \lambda_t x_t^2\right\rangle, \tag{C.44}$$

which gives

$$\frac{\mathrm{d}M_{2t}}{\mathrm{d}t} = -2\lambda_t M_{2t} + 2A_t Q_t + 2b_t M_{1t}\,, \tag{C.45}$$

where $Q_t = \langle\mu(t)\Theta[x_s(t) - x_c]x_t(t)\rangle$ represents another quantity to be computed. It can be expressed as follows

$$Q_t = e^{-\lambda_t t}\int_0^t e^{\lambda_t r}\left(A_t\left\langle\mu(r)\Theta[x_s(r) - x_c]\mu(t)\Theta[x_s(t) - x_c]\right\rangle + b_t\left\langle\mu(t)\Theta[x_s(t) - x_c]\right\rangle\right)\mathrm{d}r\,. \tag{C.46}$$

We solve the last expectation by using the previous independence assumption, whereas we can approximate the first expectation as done for $\left\langle\Theta[x_s(r) - x_c]\Theta[x_s(r') - x_c]\right\rangle$. In other words we can write

$$\left\langle\mu(r)\Theta_r\mu(r')\Theta_{r'}\right\rangle = \left[\left\langle\Theta_r\right\rangle + \left(\left\langle\Theta_r\right\rangle\left\langle\Theta_{r'}\right\rangle - \left\langle\Theta_r\right\rangle\right)\cdot\left(1 - e^{-\lambda_s(r'-r)}\right)\right]q(1,r,r')\,, \tag{C.47}$$

where we have used the short notation Θ_r and $\Theta_{r'}$ for $\Theta[x_s(r)-x_c]$ and $\Theta[x_s(r')-x_c]$, respectively. By using this approximation, in later steps we need to assume $q(1,r,r') = q(1,r)q(1,r')$. For this

reason we assume this independence at this earlier stage by using the following approximation:

$$\langle \mu(r)\Theta_r \mu(r')\Theta_{r'} \rangle = \Big[\langle \Theta_r \rangle + \Big(\langle \Theta_r \rangle \langle \Theta_{r'} \rangle - \langle \Theta_r \rangle \Big) \cdot \Big(1 - e^{-\lambda_s (r'-r)} \Big) \Big] q(1,r) q(1,r') \,. \tag{C.48}$$

Using these approximations, the ODE for Q_t becomes

$$\begin{aligned} \frac{\mathrm{d}Q_t}{\mathrm{d}t} &= -\lambda_t Q_t + (A_t q(1,t) + b_t) \langle \Theta[x_s(t) - x_c] \rangle \, q(1,t) \\ &+ M_{1t} \left[\langle \Theta[x_s(t) - x_c] \rangle \, (-g q(1,t) + g_+) + q(1,t) \frac{\mathrm{d}}{\mathrm{d}t} \langle \Theta[x_s(t) - x_c] \rangle \right] \\ &+ e^{-(\lambda_s + \lambda_t)t} A_t \left[(\langle \Theta[x_s(t) - x_c] \rangle - 1)(q(1,t)\lambda_s - g q(1,t) + g_+) - q(1,t) \frac{\mathrm{d}}{\mathrm{d}t} \langle \Theta[x_s(t) - x_c] \rangle \right] I \,, \end{aligned} \tag{C.49}$$

where the integral

$$I = \int_0^t e^{(\lambda_s + \lambda_t) r} \, \langle \Theta[x_s(r) - x_c] \rangle \, q(1,r) \mathrm{d}r \,. \tag{C.50}$$

is solved numerically.

C.2.1 Results with AND gate FFL

We report results with AND gate FFL on simulated data set, omitted in Section 4.3. Considerations are the same as for the OR gate FFL.

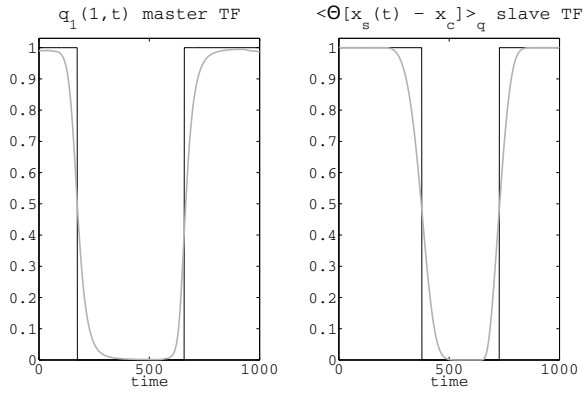

Figure C.1: Results on simulated data with AND gate FFL. Inferred activity for master and slave transcription factors (thick grey), compared with true inputs (black).

C.3 Additional implementation details

C.3.1 Initialisation of kinetic parameters

Initialisation of parameter λ_s is done by evaluating the exponential decay constant for the longest monotonic path of the slave gene data (after a moving average filter). A_s and b_s are initialised by considering that the model is bistable with $x_{s\,high} = (A_s + b_s)/\lambda_s$ as higher steady

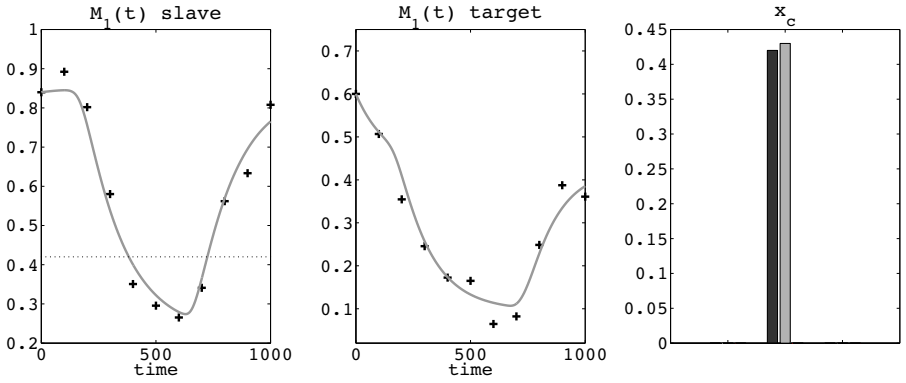

Figure C.2: Results on simulated data with AND gate FFL. Left and centre: posterior first moments (grey) versus noisy observations (crosses) for slave and target gene. Right: estimated x_c (grey), compared to true one (black).

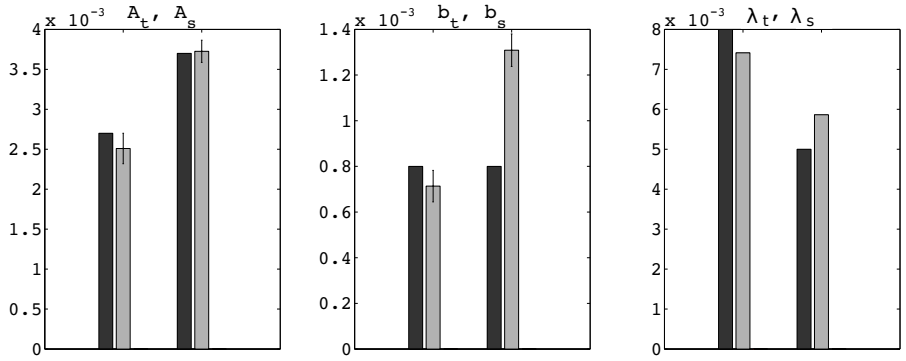

Figure C.3: Results on simulated data with AND gate FFL. Estimated parameters (grey), compared to true values (black).

state and $x_{s\,low} = b_s/\lambda_s$ as lower steady state. The same procedure is adopted to initialise λ_t, b_t and A_t (and also A_c for the OR gate FFL) using the target gene data.

C.3.2 Prior over the critical threshold

Critical threshold x_c is chosen by running different simulations with different x_c values and monitoring the final value of the variational free energy. We place a prior distribution on x_c given by a Gaussian distribution centred at half of x_s gene expression ($\bar{x}_c$) with standard deviation σ_{x_c}. By doing this, we add to the variational free energy a term

$$-\mathbb{E}_q\left[\log p_{prior}(x_c)\right] = -\log\left[\frac{1}{\sqrt{2\pi\sigma^2_{x_c}}}\exp\left(-\frac{(x_c-\bar{x}_c)^2}{2\sigma^2_{x_c}}\right)\right], \tag{C.51}$$

which does not affect the functional derivatives reported above.

C.3.3 Test on the quality of the inference

We assess the quality of the inference model by comparing the inferred posterior master TF activity $q(1,t)$ with the true master TF activity $\mu(t)$. This is be done also for the slave TF, by comparing $\langle\Theta\left[x_s(t)-x_c\right]\rangle$ with the true $\Theta\left[x_s(t)-x_c\right]$. Running 250 simulations with different model parameters, the correlation between inferred and true TF activity is around 0.94 for the master TF and 0.93 for the slave TF in the OR gate FFL, and around 0.86 for the master TF and 0.95 for the slave TF in the AND gate FFL. The lower correlation values in the AND gate FFL with respect to the OR gate FFL case are probably due to a higher level of approximation of the moments involving the Heaviside step function.

C.4 Robustness to Gamma distributed noise

In our model we assume to have observation corrupted by i.i.d. zero mean Gaussian noise. This assumption is not properly correct, so we tested the inference model using a non-Gaussian noise source: a Gamma distributed noise $X \sim \Gamma(\kappa,\theta)$. This distribution has a skewness regulated by the parameter κ. Parameters κ and θ are set such that the variance of the distribution, $\kappa\theta^2$, is comparable with the variance of the Gaussian noise we have previously used.

Results of the inference (Fig. C.4) are compared by computing the correlation between the inferred TF activity and the real TF activity, for both master TF and slave TF. For the master TF activity, we obtain same values of correlation both with Gaussian and Gamma noise; for the slave TF activity we obtain a non significative lower correlation using Gamma noise (difference of $\sim 0.01 - 0.02$). These results show that the inference quality is not affected by the different source of noise and the assumption of Gaussian noise corrupted observations is correct in practice.

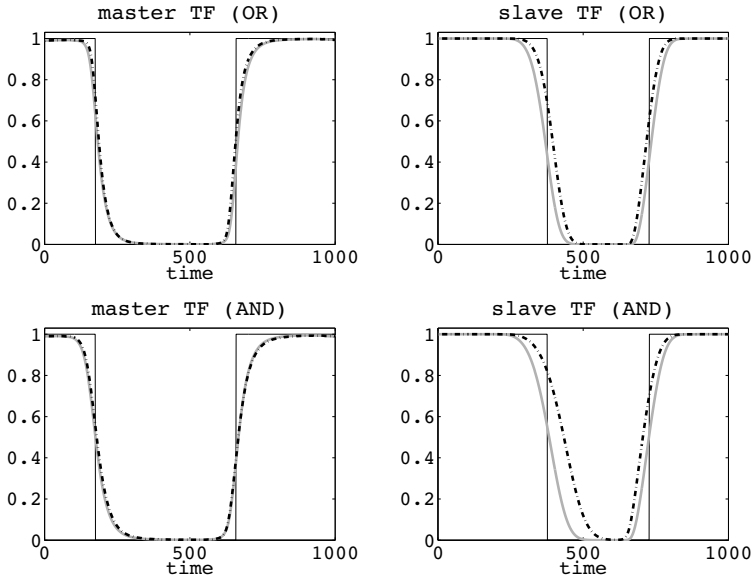

Figure C.4: Inferred TF activities using Gaussian noise (solid grey) and gamma noise (dashed black) compared to true profiles (solid black). Top: OR gate FFL; bottom: AND gate FFL. Gaussian noise has zero mean and standard deviation 0.03; gamma noise has parameters $\kappa = 2.56$ and $\theta = 0.02$, with standard deviation of 0.03. Parameters for data simulations with OR gate FFL are: $A_s = 3.7 \cdot 10^{-3}$, $b_s = 0.8 \cdot 10^{-3}$, $\lambda_s = 5 \cdot 10^{-3}$, $A_t = 2.7 \cdot 10^{-3}$, $b_t = 0.8 \cdot 10^{-3}$, $\lambda_t = 8 \cdot 10^{-3}$, $A_c = 2.5 \cdot 10^{-3}$. Parameters for data simulations with AND gate FFL are: $A_s = 3.7 \cdot 10^{-3}$, $b_s = 0.8 \cdot 10^{-3}$, $\lambda_s = 5 \cdot 10^{-3}$, $A_c = 2.7 \cdot 10^{-3}$, $b_t = 0.8 \cdot 10^{-3}$, $\lambda_t = 8 \cdot 10^{-3}$.

C.5 Experimental platform for *p53* data set

C.5.1 Inference of *p53* activity using a SIM network motif

Prediction of $p53$ activity in a SIM model is done by using mRNA time courses of $p53$ target genes ($DDB2$, $p21$, BIK). These genes are part of the original target genes used by Barenco and colleagues (Barenco et al., 2006). In addition they are regulated by $E2F1$ as well, therefore they can be used also for the FFL model. We have used also other target genes ($PUMA$, $SIVA$ and $DRAM$) giving a SIM model as showed in Figure C.5.

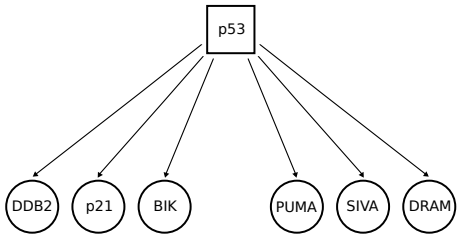

Figure C.5: SIM network motif involving $p53$ and its target genes $DDB2$, $p21$, BIK, $PUMA$, $SIVA$ and $DRAM$.

Prediction with the SIM model was compared to Barenco *et al.*'s prediction and to nor-

malised experimental p53 activity[3] (Barenco et al., 2006). These results are reported in Figure 4.10. This comparison shows that our SIM model, as well as Barenco *et al.*'s model, does not predict the decrease of the p53 activity at the experimentally measured 10 h time.

C.5.2 Inference of *p53* activity using a FFL network motif

We built a framework to infer the activity of the master TF ($E2F1$) and slave TF ($p53$) using mRNA time-courses data from the following genes:

- genes combinatorially regulated by $p53$ and $E2F1$
  y_{FFL} = { $DDB2$, $p21$, BIK, $PUMA$, $SIVA$, $DRAM$ }
- genes regulated only by $E2F1$
  y_{SIM} = { $MCM7$, $LIG1$, $MCM5$ }.

The structure of this gene network represents a mixture of two subnetworks: a SIM network motif and a FFL network motif, as depicted in Section 4.4. To be precise, the FFL is a multi-output FFL since we are considering a FFL with more than one target gene.

The inference procedure is composed of a pre-optimisation part, focusing only on the SIM portion of the network, followed by a global optimisation part, regarding the whole network. During the pre-optimisation, y_{SIM} genes together with $p53$ are used in a SIM model to infer the activity of $E2F1$. In this part we also estimate parameters for genes y_{SIM} and $p53$ (all regulated by $E2F1$).

The inferred activity of $E2F1$ and estimated parameters are then used to initialise $E2F1$ activity and model parameters in the global optimisation part. In addition to the SIM part, here we consider also the FFL part. We use mRNA time-courses of all the genes (y_{SIM} and y_{FFL}) to infer the activity of both $E2F1$ and $p53$. The optimisation algorithm is similar to that for a simple FFL network. The main difference is the computation of additional moments (for genes y_{SIM}), additional Lagrange multipliers and slightly different gradients for the transition rates.

Microarray time-courses are obtained from a freely available data set produced by Barenco and colleagues (Barenco et al., 2009). They contain three independent replicates for each gene; each replicate represents a gene expression profile at times $[0, 2, 4, 6, 8, 10, 12]$ h. For each gene of interest we normalise the mean expression profile of the three replicates.

The predicted activity of $p53$ is showed in Figure 4.10. Here we report the inferred activity of $E2F1$ and first moments for all genes, after the global optimisation (Fig. C.6 and C.7). Genes regulated solely by $E2F1$ (SIM part) have been included in the optimisation procedure only for reason of completeness. Removing the SIM part of the model (and considering only the FFL part) does not affect the results. Figure C.8 shows results obtained by removing the SIM part during the global optimisation, and considering only a FFL with $E2F1$, $p53$ and $DRAM$.

[3]Note that usually an experimental TF activity is not available, but only the mRNA time courses of TF's target genes.

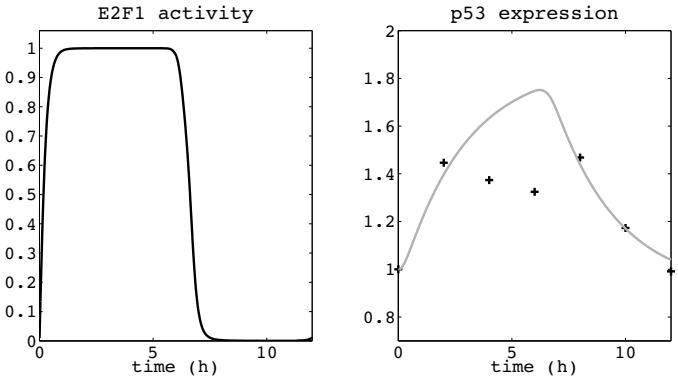

Figure C.6: Left: posterior inferred $E2F1$ activity. Right: posterior first moment of $p53$ mRNA expression (solid grey) compared to observations (crosses).

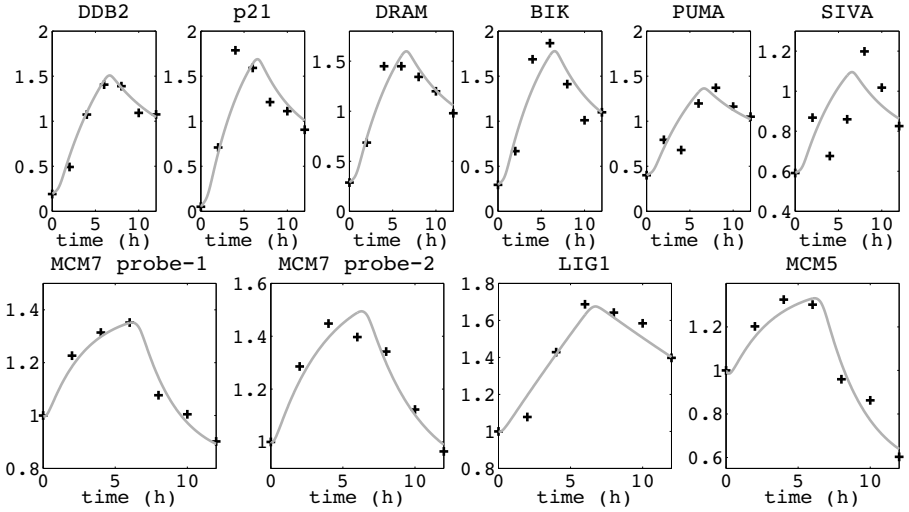

Figure C.7: Top: posterior first moments of FFL regulated targets (solid) compared to observations (crosses). Bottom: posterior first moments of targets regulated solely by $E2F1$ (solid) compared to observations (crosses).

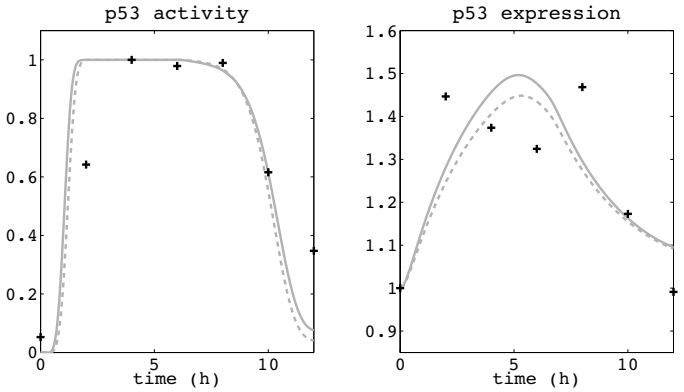

Figure C.8: Results obtained using the FFL $E2F1$-$p53$-$DRAM$. Left: posterior $p53$ activity inferred using AND gate FFL (solid) and OR gate FFL (dashed), compared to experimental measurements (crosses). Right: posterior first moment of p53 mRNA expression obtained with AND gate FFL (solid) and OR gate FFL (dashed), compared to observations (crosses).

C.6 Experimental platform for *E. coli* data set

For the *E. coli* data set we used the same experimental set up described for the $p53$ data set: first we used genes solely regulated by CRP ($ppdD$, agp and mlc) in a pre-optimisation part. Then we used also the target gene $manX$ involved in OR gate FFL with CRP and mlc.

Bacterial data, obtained by Partridge and colleagues (Partridge et al., 2007), consist of 5 points mRNA time-courses at $[0, 5, 10, 15, 60]$ min. In Figure C.9 we report additional inference results: prediction of the mlc activity and first moments for genes $ppdD$ and agp. Expression levels of CRP mRNA are also reported in Figure C.9 (top right). This behaviour in the CRP gene expression could also suggest a further explanation for the peak in the CRP activity (Fig. 4.11). A mechanism of negative autoregulation (Hanamura and Aiba, 1991), which we are not taking into account in the present work, could be postulated.

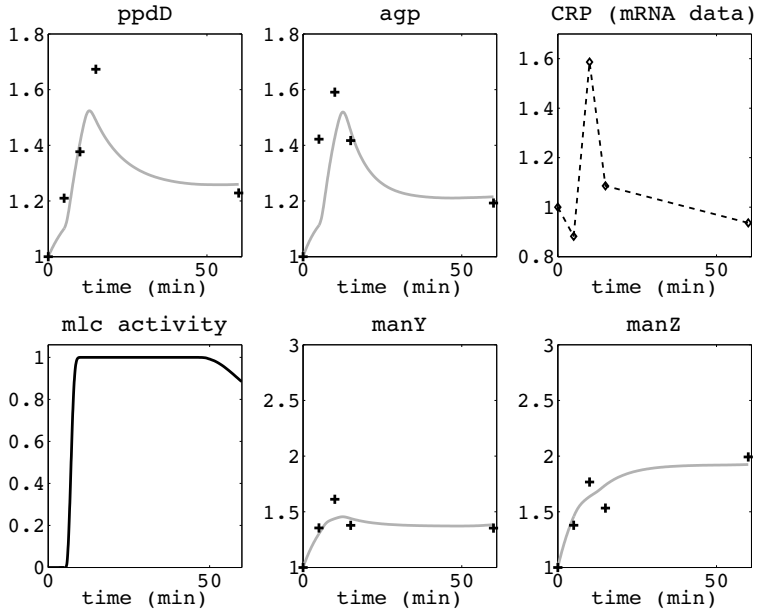

Figure C.9: Results on *E. coli* data set with FFL. Top left and centre: posterior first moment (solid) of $ppdD$ and agp, compared to observations (crosses). Top right: CRP expression levels (diamonds). Bottom left: posterior inferred Mlc activity using a combined SIM-FFL model. Bottom centre and right: posterior first moment (solid) of $manY$ and $manZ$ (right), compared to observations (crosses).

C.6.1 Inference of *CRP* activity in a SIM network motif

We report results of the activity of CRP using a SIM model, where the master regulator CRP controls the expression of target genes mlc and $manXYZ$. It is interesting to compare these results (Fig. C.10) with results obtained using the FFL model (Fig. 4.11 in Section 4.5). It is clear how the posterior activity of CRP is different in the two cases: a restriction of the inference model to SIM, prevents the CRP activity to have a proper peak. The difference in the inferred kinetic parameters between the two cases is presumably due to the lack of the

interaction between mlc and $manX$ in the SIM model.

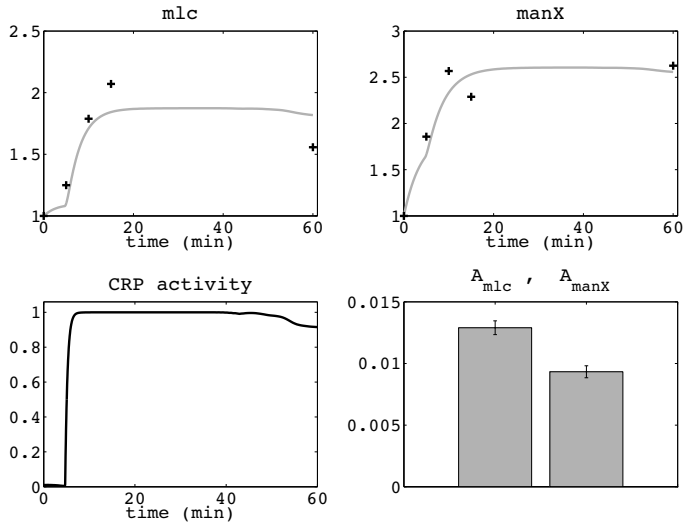

Figure C.10: Results on *E. coli* data set with SIM. Top: posterior inferred mean of $manX$ and mlc (solid) compared to observations (crosses). Bottom left: posterior CRP activity. Bottom right: estimated parameters of the SIM architecture.

C.7 Laplace approximation

The Laplace approximation is a simple method to approximate an intractable density

$$p(\boldsymbol{X}) = \frac{1}{\mathcal{Z}} f(\boldsymbol{X}) \,,$$

with a Gaussian density. It is obtained by Taylor expanding the logarithm of the unnormalised probability $f(\boldsymbol{X})$ around the maximum (mode) X_0 of the distribution:

$$\log f(\boldsymbol{X}) \simeq \log f(\boldsymbol{X}_0) + (\boldsymbol{X} - \boldsymbol{X}_0)^{\mathrm{T}} \nabla \log f(\boldsymbol{X}) \mid_{\boldsymbol{X}=\boldsymbol{X}_0} + \frac{1}{2}(\boldsymbol{X} - \boldsymbol{X}_0)^{\mathrm{T}} \mathrm{H} (\boldsymbol{X} - \boldsymbol{X}_0) \,.$$

The mode X_0 is found by setting to zero the gradient $\nabla f(\boldsymbol{X}) \mid_{\boldsymbol{X}=\boldsymbol{X}_0}$. H is the Hessian matrix, which is defined by $\mathrm{H} = \nabla\nabla \log f(\boldsymbol{X}) \mid_{\boldsymbol{X}=\boldsymbol{X}_0}$. Since $\boldsymbol{X}_0$ is the maximum of the distribution, the gradient $\nabla \log f(\boldsymbol{X}) \mid_{\boldsymbol{X}=\boldsymbol{X}_0}$ is null[4]. Then the distribution is approximated as

$$f(\boldsymbol{X}) \simeq f(\boldsymbol{X}_0) \exp\left\{ -\frac{1}{2}(\boldsymbol{X} - \boldsymbol{X}_0)^{\mathrm{T}} \bar{\mathrm{H}}^{-1} (\boldsymbol{X} - \boldsymbol{X}_0) \right\} ,$$

where $\bar{\mathrm{H}} = -\mathrm{H}^{-1}$. This can be simply normalised to become the desired Gaussian density

$$\mathcal{N}\left(\boldsymbol{X} | \boldsymbol{X}_0, \bar{\mathrm{H}}\right) = \frac{1}{(2\pi)^{D/2}} \frac{1}{\left|\bar{\mathrm{H}}\right|^{1/2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{X} - \boldsymbol{X}_0)^{\mathrm{T}} \bar{\mathrm{H}}^{-1} (\boldsymbol{X} - \boldsymbol{X}_0) \right\} ,$$

[4]In the simpler univariate case: $\left.\frac{\mathrm{d}\log f(X)}{\mathrm{d}X}\right|_{X=X_0} = \left.\frac{\mathrm{d}\log f(X)}{\mathrm{d}f(X)} \frac{\mathrm{d}f(X)}{\mathrm{d}X}\right|_{X=X_0}$, where $\left.\frac{\mathrm{d}f(X)}{\mathrm{d}X}\right|_{X=X_0} = 0$.

with D the dimension of vector $\boldsymbol{X}$. Once we have approximated our distribution with a Gaussian, it is straightforward to compute the evidence $\mathcal{Z}$ or expectations of interest. We have assumed a unique mode $\boldsymbol{X}_0$, but multimodality is common in multivariate distributions. In this case we have different approximations for different modes (Bishop, 2006).

Appendix D

Appendix to Chapter 5

D.1 Modelling light input

During inference and predictions, we model a light input that affects the $TOC1$ transcription factor. This is done by adding a contribution to the switching rates of the promoter of $TOC1$'s target gene. The target gene is $CCA1$ in the NFL model and X in the $TOC1$-X-$CCA1$ repressilator model. E.g. the new switching rates in the repressilator $TOC1$-X-$CCA1$ become:

$$f'_{X+} = f_{X+} + f_{+}(I)\,, \tag{D.1}$$

$$f'_{X-} = f_{X-} + f_{-}(I)\,, \tag{D.2}$$

where $f_{X\pm}$ represent the contribution of the upstream protein to the switching rates. The terms $f_{\pm}(I)$ represent the light contribution, which is a square wave proportional to the light input. Ideally, it should be possible to infer the amplitude of the square wave $f_{\pm}(I)$ from the data; however, in our experiments on simulated data we found that this did not give reliable results. We therefore fixed amplitude of the light contribution in this fashion: we run the inference on a L:D 12:12 without modelling the light input. We then assumed the amplitude of the light input to be a small fraction (0.05) of the average posterior rate during the L periods[1]. The model's predictions were robust for variations of the amplitude of the light input within a range $0.01 - 0.1$.

D.2 Additional results

D.2.1 Results with repressilator *TOC1-X-CCA1*

We report additional results of the inference from transcriptional/translational reporters and additional predictions on altered photoperiod data. Figure D.1 (left) shows the fitting of the repressilator $TOC1$-X-$CCA1$ to the translational reporter data.

Figure D.2 shows prediction results on data where a single (altered) dark period is followed by constant light. Both repressilator $TOC1$-X-$CCA1$ and NFL model predict accurately a stable

[1] We use a fraction of the average of the posterior f_{X+} during the L periods to model the amplitude of $f_{+}(I)$, and the same fraction of the average of the posterior f_{X-} during the L periods to model the amplitude of $f_{-}(I)$.

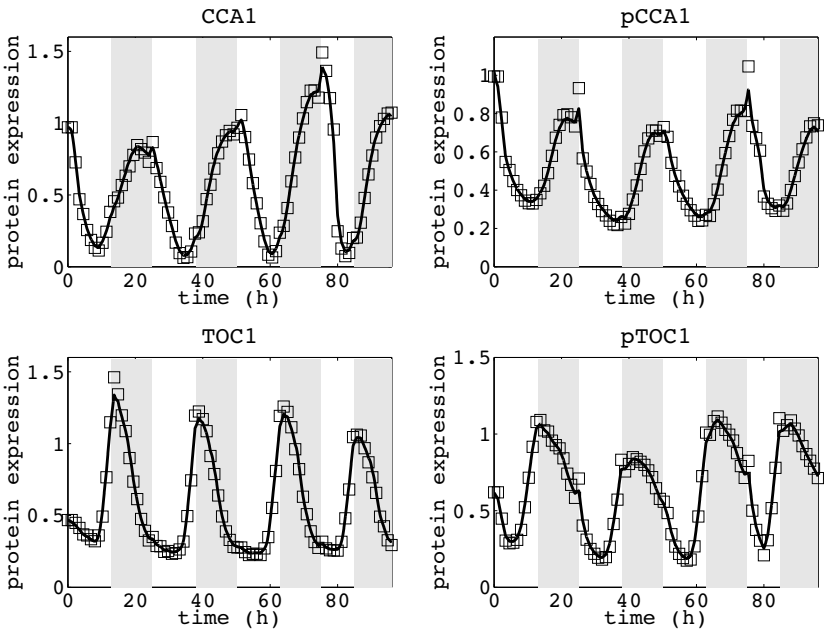

Figure D.1: Left: posterior mean protein states (solid lines) compared to luminescence translational reporter data (empty squares), using the hybrid repressilator $TOC1$-X-$CCA1$ model. Right: posterior mean protein states (solid lines) compared to luminescence transcriptional reporter data (empty squares), using the model in Figure 5.9B.

phase shift induced by dark periods of different lengths, as well as a dampening oscillatory behaviour during constant light.

D.2.2 Robustness to initial parameter values

In contrast to stochastic optimisation methods such as MCMC, variational methods are deterministic. Therefore, the minimum reached is only one and depends on initial conditions; by changing the initialisation, a different minimum is found. The standard procedure consists in using different initialisations. Figure D.3 shows the variation of the parameter estimates, when changing the initial conditions. Results of fitting to the protein data are quite robust to the different set of parameters, as latent processes can adapt very well to the small changes in parameter values.

D.2.3 Results with repressilator *CCA1-X-TOC1*

We report results obtained with the repressilator structure $CCA1$-X-$TOC1$. Figure D.4 shows the inferred promoter states for $CCA1$ (left) and $TOC1$ (center), obtained using translational (solid lines) and transcriptional (dashed lines) reporters. Figure D.4 (right) shows the mean prediction of the hypothetical gene X. This repressilator structure fails to predict the promoter states, showing that a structure where the hypothetical gene X is repressed by $CCA1$ is not consistent with the data.

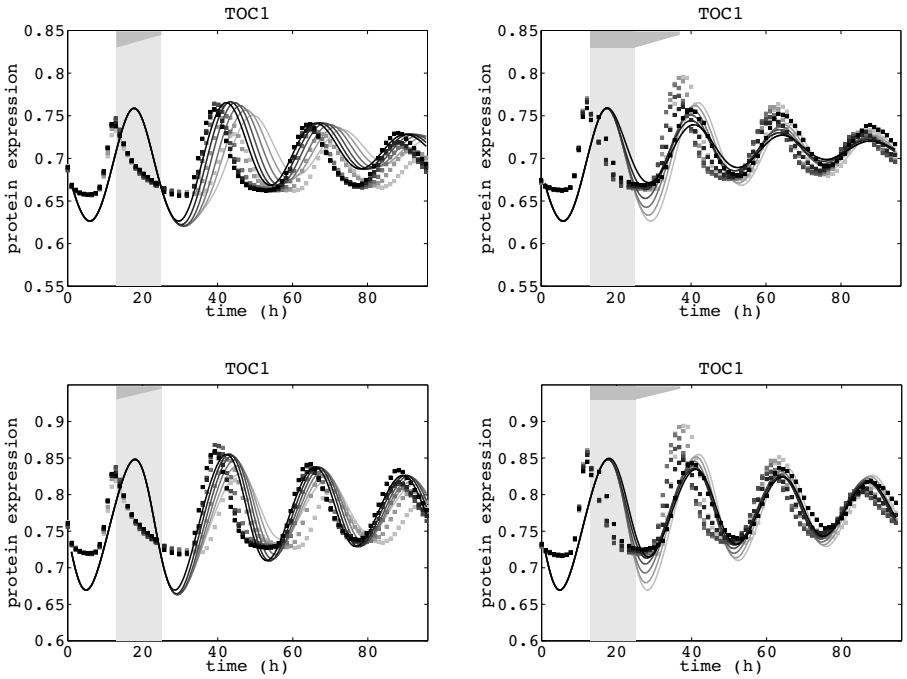

Figure D.2: Prediction in constant light. Upper panel: hybrid repressilator $TOC1$-X-$CCA1$ model results. Bottom panel: hybrid NFL model results. Prediction of $TOC1$ protein level (solid lines) compared to observations (squares) in constant light, after dark periods of different lengths. Colors from darker to lighter are obtained with progressive increments (2h increment) in the length of dark period. Left: darkest line (dark period length of 2h), lightest line (dark period length of 12h). Right: darkest line (dark period length of 12h), lightest line (dark period length of 22h).

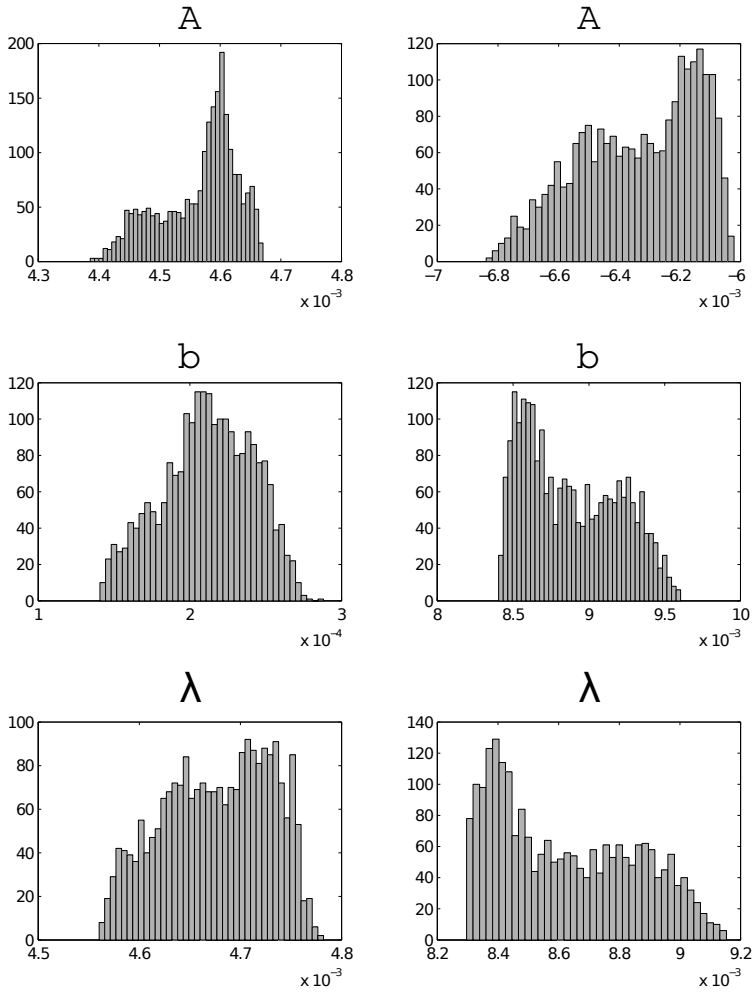

Figure D.3: Histograms showing the distribution of estimated parameters in the negative feedback loop model, using different initial conditions.

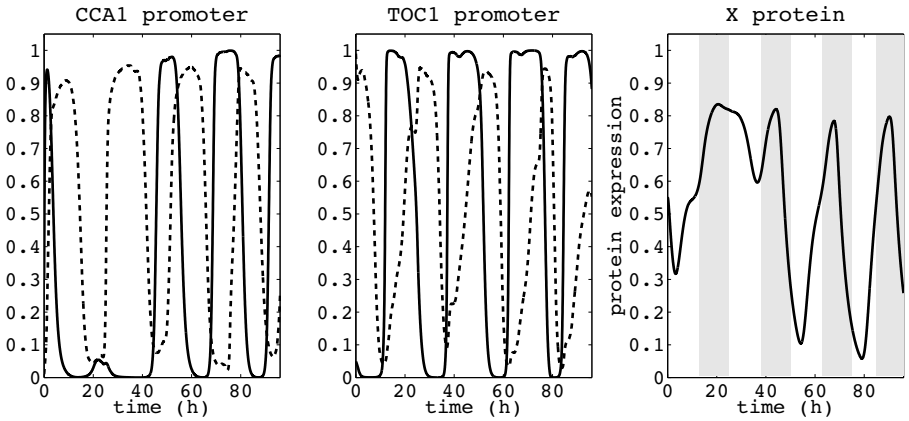

Figure D.4: Inferred promoter states for $CCA1$ (left) and $TOC1$ (center). Results obtained with the hybrid repressilator model $CCA1$-X-$TOC1$ (solid lines) using translational reporters; results obtained with the model in Figure 5.9B (dashed lines) using transcriptional reporters. Right: mean prediction of the hypothetical gene X.

D.3 Stochastic optimisation of ODE models

We report the optimisation used to find the parameters in Cantone *et al.*'s delay differential equation (DDE) model (Cantone et al., 2009). The same method has been used to optimise the parameters in Troein *et al.*'s ODE model (Troein et al., 2011).

Cantone *et al.*'s model consists of the following five nonlinear DDEs:

$$\frac{\mathrm{d}x_1}{\mathrm{d}t} = \alpha_1 + v_1 \left(\frac{x_3^{h_1}(t-\tau)}{(k_1^{h_1} + x_3^{h_1}(t-\tau)) \left(1 + \frac{x_5^{h_2}}{k_2^{h_2}}\right)} \right) - d_1 x_1 \,, \tag{D.3}$$

$$\frac{\mathrm{d}x_2}{\mathrm{d}t} = \alpha_2 + v_2 \left(\frac{x_1^{h_3}}{k_3^{h_3} + x_1^{h_3}} \right) - (d_2 - \Delta(\beta_1)) x_2 \,, \tag{D.4}$$

$$\frac{\mathrm{d}x_3}{\mathrm{d}t} = \alpha_3 + v_3 \left(\frac{x_2^{h_4}}{k_4^{h_4} + x_2^{h_4} \left(1 + \frac{x_4^4}{\gamma^4}\right)} \right) - d_3 x_3 \,, \tag{D.5}$$

$$\frac{\mathrm{d}x_4}{\mathrm{d}t} = \alpha_4 + v_4 \left(\frac{x_1^{h_5}}{k_5^{h_5} + x_1^{h_5}} \right) - (d_4 - \Delta(\beta_2)) x_4 \,, \tag{D.6}$$

$$\frac{\mathrm{d}x_5}{\mathrm{d}t} = \alpha_5 + v_5 \left(\frac{x_1^{h_6}}{k_6^{h_6} + x_1^{h_6}} \right) - d_5 x_5 \,, \tag{D.7}$$

$$\tag{D.8}$$

where $\Delta(\beta_i)$ is a rectangular window of length 10 min with amplitude β_i; τ represents a delay of $\tau = 100$ min. We learn all parameters of the model (30 in total)[2], by using a standard Metropolis Markov chain Monte Carlo (MCMC) method.

We sample a new guess of a parameter $\theta^\star$ from a Gaussian proposal distribution $\mathcal{N}(\theta^\star|\theta^t, \sigma)$, which depends on the current state θ^t. The variance σ of the proposal distribution is set in order to ensure an acceptance rate in the range $30 - 40\%$. The proposal distributions for all the parameters are all Gaussian with different values for the variances. Hill coefficients h_i (with $i = [1, \ldots, 6]$) are sampled uniformly from integers between 1 and 10.

After drawing a new sample $\theta_i^\star$ for each of the thirty parameters ($i = [1, \ldots, 30]$), we use them to solve the DDE and produce estimated trajectories. Then we compute a Gaussian likelihood $\mathcal{L}(\boldsymbol{\theta^\star})$ between data from the switch-on transition and discrete points from these trajectories. The samples $\boldsymbol{\theta^\star}$ are accepted with probability $g = \min(1, \alpha)$, where

$$\alpha = \frac{\mathcal{L}(\boldsymbol{\theta^\star})}{\mathcal{L}(\boldsymbol{\theta^t})} \,, \tag{D.9}$$

and $\mathcal{L}(\boldsymbol{\theta^t})$ represents a Gaussian likelihood obtained using the parameters $\boldsymbol{\theta^t}$. If the sample is accepted, then it becomes the current state, $\boldsymbol{\theta^{t+1}} = \boldsymbol{\theta^\star}$, otherwise the state is not updated, $\boldsymbol{\theta^{t+1}} = \boldsymbol{\theta^t}$.

In practice, at each iteration we compute α and then we draw a sample u from a uniform distribution with support $[0, 1]$. We accept the new samples $\boldsymbol{\theta^\star}$ if $\alpha > u$.

If we want to run 2×10^5 MCMC iterations for $N = 10$ different initialisations, we need

[2]The length of the delay τ is fixed to the value from the original paper ($\tau = 100\,\mathrm{min}$).

to solve the systems of DDE $2 \times 10^5 \times N$ times. This took us about 7 hours on a standard desktop machine (Intel Core 2 Duo at 3GHz). It is possible to reduce the running time by using approximate methods based on gradient matching (Calderhead et al., 2009).

Once we have estimated all the parameters, we use them to simulate the DDEs for the switch-off transition. This is done by setting the amplitudes of the rectangular windows to zero ($\beta_{1,2} = 0$) and using initial conditions for the switch-off transition, as reported in (Cantone et al., 2009). Cantone and colleagues use different values for the parameters v_3, k_4 and γ_4 during the switch-on and switch-off transitions, but we cannot do that since our aim is to use only the switch-on transition data to train the DDE model.

D.4 Calculations for approximate inference method

D.4.1 Expectation of exponential term

We are interested in the value of $\langle e^x \rangle_{q_x}$, which we can compute analytically if q_x is Gaussian. If q_x is not Gaussian, we can compute it using a Laplace approximation. Assuming $q_x = \mathcal{N}(x|m, c^2)$ we can write

$$\begin{aligned} \langle e^x \rangle_{q_x} &= \int \frac{1}{\sqrt{2\pi c^2}} e^{-\frac{(x-m)^2}{2c^2}} e^x \mathrm{d}x = \int \frac{1}{\sqrt{2\pi c^2}} e^{-\frac{x^2+m^2-2mx-2c^2x}{2c^2}} \mathrm{d}x \\ & \int \frac{1}{\sqrt{2\pi c^2}} e^{-\frac{x^2-2(c^2+m)x+(m^2+2mc^2+c^4)-2mc^2-c^4}{2c^2}} \mathrm{d}x \,, \end{aligned} \tag{D.10}$$

which finally gives

$$\langle e^x \rangle_{q_x} = \int \frac{1}{\sqrt{2\pi c^2}} e^{-\frac{(x-(m+c^2))^2}{2c^2}} e^{m+\frac{1}{2}c^2} \mathrm{d}x = e^{m+\frac{1}{2}c^2} \,. \tag{D.11}$$

Therefore, with simply obtain

$$\langle a \exp(bx) \rangle_{q_x} = a \langle \exp(bx) \rangle_{q_x} = a \exp\left(b\, m + \frac{1}{2} c^2 b^2 \right) . \tag{D.12}$$

D.4.2 Update formula for posterior switching rates

We derive the update formula for the posterior switching rates. By setting to zero the functional derivative of the Lagrangian

$$\begin{aligned} \mathcal{L}[q_\mu, g_\pm, \psi] &= \langle KL\left[q_\nu(\nu) \| p(\nu|\chi') \right] \rangle_{q_{\chi'}} + \int \mathrm{d}t \frac{1}{2\sigma^2} [A^2 - 2A(\alpha+\lambda)m - 2A(\beta - b)] q_\mu(1,t) \\ &+ \int \mathrm{d}t\, \psi(t) \left(\frac{\mathrm{d}q_\mu(1,t)}{\mathrm{d}t} + (g_- + g_+) q_\mu(1,t) - g_+ \right) , \end{aligned} \tag{D.13}$$

with respect to g_+, we obtain

$$(1 - q_\mu(1,t)) \frac{\delta}{\delta g_+} \Big(g_+ \log(g_+) - g_+ \langle \log(f_+) \rangle_{q_{\chi'}} - g_+ \Big) + \psi q_\mu(1,t) - \psi = 0 \,.$$

From this expression we obtain

$$(1 - q_\mu(1,t)) \left(\log(g_+) - \langle \log(f_+) \rangle_{q_{\chi'}} \right) - (1 - q_\mu(1,t))\psi = 0\,,$$

which gives the update formula for g_+. The same derivation is done for g_-.

D.4.3 Equation for the *r* variable

We derive the equation for the variable r from the following equation for the Lagrange multiplier ψ

$$\begin{aligned} \frac{\mathrm{d}\psi}{\mathrm{d}t} &= \left[g_- \left\langle \log \frac{g_-}{f_-} \right\rangle_{q_{\chi'}} + \langle f_- \rangle_{q_{\chi'}} - g_- \right] - \left[g_+ \left\langle \log \frac{g_+}{f_+} \right\rangle_{q_{\chi'}} + \langle f_+ \rangle_{q_{\chi'}} - g_+ \right] \\ &+ (g_- + g_+)\psi + \frac{1}{2\sigma^2}[A^2 - 2A(\alpha + \lambda)m - 2A(\beta - b)]\,. \end{aligned} \tag{D.14}$$

By using the update formula for the posterior switching rates ($\log g_\pm = \langle \log f_\pm \rangle_{q_{\chi'}} \pm \psi$), the ODE for the Lagrange multiplier reduces to

$$\frac{\mathrm{d}\psi}{\mathrm{d}t} = \left(\langle f_- \rangle_{q_{\chi'}} - g_- \right) - \left(\langle f_+ \rangle_{q_{\chi'}} - g_+ \right) + \frac{1}{2\sigma^2}[A^2 - 2A(\alpha + \lambda)m - 2A(\beta - b)]\,. \tag{D.15}$$

Now, by setting $r = \exp(-\psi)$ and substituting the posterior switching rates with

$$\begin{aligned} g_+ &= k_p \, \exp(k_e m') \, r^{-1}\,, \\ g_- &= k_m \, r\ , \end{aligned}$$

we obtain the ODE for the variable r

$$\frac{\mathrm{d}r}{\mathrm{d}t} = k_p \exp(k_m m') \Big[r \, \exp\left(\frac{1}{2} c'^2 k_e^2 \right) - 1 \Big] - k_m \, r(1-r) - \frac{1}{2\sigma^2}[A^2 - 2A(\alpha+\lambda)m - 2A(\beta-b)]r\,, \tag{D.16}$$

where we have used the following expectations

$$\begin{aligned} \langle f_+ \rangle_{q_{\chi'}} &= k_p \exp\left(k_e m' + \frac{c'^2 k_e^2}{2} \right), \\ \langle f_- \rangle_{q_{\chi'}} &= k_m\,, \end{aligned}$$

and

$$\frac{\mathrm{d}\psi}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}(-\log r) = -\frac{1}{r}\frac{\mathrm{d}r}{\mathrm{d}t}\,. \tag{D.17}$$

Bibliography

Acar, M., Mettetal, J. T., and van Oudenaarden, A. (2008). Stochastic switching as a survival strategy in fluctuating environments. *Nature Genetics*, 40(4):471–5.

Alon, U. (2006). *An introduction to systems biology: design principles of biological circuits*. Champan and Hall, London.

Archambeau, C., Cornford, D., Opper, M., and Shawe-Taylor, J. (2007). Gaussian process approximations of stochastic differential equations. *Journal of machine learning research*, 1:1–16.

Asif, H. M. S. and Sanguinetti, G. (2011). Large scale learning of combinatorial transcriptional dynamics from gene expression. *Bioinformatics*, 27(9):1277–1283.

Barber, D. (2012). *Bayesian Reasoning and Machine Learning*. Cambridge University Press, Cambridge.

Barenco, M., Papouli, E., Shah, S., Brewer, D., Miller, C., and Hubank, M. (2009). rHVDM: an R package to predict the activity and targets of a transcription factor. *Bioinformatics*, 25(3):419–420.

Barenco, M., Tomescu, D., Brewer, D., Callard, R., Stark, J., and Hubank, M. (2006). Ranked prediction of p53 targets using hidden variable dynamic modeling. *Genome biology*, 7(3):R25.

Barnes, C. P., Silk, D., Sheng, X., and Stumpf, M. P. H. (2011). Bayesian design of synthetic biological systems. *Proceedings of the National Academy of Sciences*, 108(37):15190–5.

Berg, J., Tymoczko, J., and Stryer, L. (2002). *Biochemistry*. WH Freeman, New York.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer, New York.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer, New York.

Boys, R. J., Wilkinson, D. J., and Kirkwood, T. B. (2008). Bayesian inference for a discretely observed stochastic kinetic model. *Statistics and Computing*, 18(2):125–135.

Bracken, A., Ciro, M., Cocito, A., and Helin, K. (2004). E2F target genes: unraveling the biology. *Trends in biochemical sciences*, 29(8):409–417.

Brunel, N. and d'Alché Buc, F. (2010). Flow-based bayesian estimation of nonlinear differential equations for modeling biological networks. In Dijkstra, T., Tsivtsivadze, E., Marchiori, E., and Heskes, T., editors, *Pattern Recognition in Bioinformatics*, pages 443–454.

Calderhead, B., Girolami, M., and Lawrence, N. D. (2009). Accelerating bayesian inference over nonlinear differential equations with gaussian processes. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 21*, pages 217–224.

Cantone, I., Marucci, L., Iorio, F., Belcastro, M. R. V., Bansal, M., Santini, S., di Bernardo, M., di Bernardo, D., and Cosma, M. (2009). A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell*, 137(1):172–181.

Chalancon, G., Ravarani, C. N., Balaji, S., Martinez-Arias, A., Aravind, L., Jothi, R., and Babu, M. M. (2012). Interplay between gene expression noise and regulatory network architecture. *Trends in Genetics*, 28(5):221–32.

Chinnadurai, G., Vijayalingam, S., and Rashmi, R. (2009). BIK, the founding member of the BH3-only family proteins: mechanisms of cell death and role in cancer and pathogenic processes. *Oncogene*, 27:S20–S29.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 39(1):1–38.

Dondelinger, F., Husmeier, D., Rogers, S., and Filippone, M. (2013). Ode parameter inference using adaptive gradient matching with gaussian processes. In Carvalho, C. M. and Ravikumar, P., editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pages 216–228.

Doucet, A. and Johansen, A. (2009). A tutorial on particle filtering and smoothing: fifteen years later. In Crisan, D. and Rozovskii, B., editors, *Handbook of Nonlinear Filtering*. Oxford University Press, Oxford.

Eldar, A. and Elowitz, M. B. (2010). Functional roles for noise in genetic circuits. *Nature*, 467(7312):167–73.

Elf, J. and Ehrenberg, M. (2003). Fast evaluation of fluctuations in biochemical networks with the linear noise approximation. *Genome Research*, 13(11):2475–2484.

Elowitz, M. and Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators. *Nature*, 403:335–8.

Elowitz, M., Levine, A., Siggia, E., and Swain, P. (2002). Stochastic gene expression in a single cell. *Science*, 297:1183–6.

Eyink, G. L., Restrepo, J. L., and Alexander, F. J. (2004). A mean field approximation in data assimilation for nonlinear dynamics. *Physica D*, 194:347–368.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3–4):601–20.

Gammaitoni, L., Hänggi, P., Jung, P., and Marchesoni, F. (1998). Stochastic resonance. *Reviews of Modern Physics*, 70(1):223–287.

Gao, P., Honkela, A., Rattray, M., and Lawrence, N. (2008). Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities. *Bioinformatics*, 24(16):i70–i75.

Gardiner, C. W. (2009). *Stochastic methods: a handbook for the natural and social sciences*, volume 4. Springer, Berlin.

Gendron, J., Pruneda-Paz, J., Doherty, C., Gross, A., Kang, S., and Kay, S. (2012). Arabidopsis circadian clock protein, toc1, is a dna-binding transcription factor. *Proceedings of the National Academy of Sciences*, 109(8):3167–3172.

Georgoulas, A., Clark, A., Ocone, A., Gilmore, S., and Sanguinetti, G. (2012). A subsystems approach for parameter estimation of ode models of hybrid systems. In Bartocci, E. and Bortolussi, L., editors, *Hybrid Systems and Biology, EPTCS 92*.

Ghahramani, Z. and Hinton, G. (1996). Parameter estimation for linear dynamical systems. *Technical Report*, pages CRG–TR–96–2.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81(25):2340–2361.

Gillespie, D. T. (2000). The chemical langevin equation. *The Journal of Chemical Physics*, 113(1):297–306.

Girolami, M. and Calderhead, B. (2011). Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214.

Golightly, A. and Wilkinson, D. J. (2005). Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics*, 61(3):781–788.

Golightly, A. and Wilkinson, D. J. (2006). Bayesian sequential inference for stochastic kinetic biochemical network models. *Journal of Computational Biology*, 13(3):838–51.

Golightly, A. and Wilkinson, D. J. (2011). Bayesian parameter inference for stochastic biochemical network models using particle markov chain monte carlo. *Interface Focus*, 1(6):807–820.

Görke, B. and Stülke, J. (2008). Carbon catabolite repression in bacteria: many ways to make the most out of nutrients. *Nature Review Microbiology*, 6(8):613–624.

Grima, R. (2010). An effective rate equation approach to reaction kinetics in small volumes: theory and application to biochemical reactions in nonequilibrium steady-state conditions. *The Journal of Chemical Physics*, 133:035101.

Grima, R. (2012). A study of the accuracy of moment-closure approximations for stochastic chemical kinetics. *The Journal of Chemical Physics*, 136:154105.

Grima, R., Schmidt, D. R., and Newman, T. J. (2012). Steady-state fluctuations of a genetic feedback loop: an exact solution. *The Journal of Chemical Physics*, 137:035104.

Grima, R., Thomas, P., and Straube, A. (2011). How accurate are the non-linear chemical fokker-planck and chemical langevin equations? *The Journal of Chemical Physics*, 135:084103.

Hanamura, A. and Aiba, H. (1991). Molecular mechanism of negative autoregulation of escherichia coli crp gene. *Nucleic Acids Research*, 19(16):4413–4419.

Haury, A., Mordelet, F., Vera-Licona, P., and Vert, J. (2012). Tigress: Trustful inference of gene regulation using stability selection. *BMC Systems Biology*, 6(1).

Heijde, M., Zabulon, G., Corellou, F., Ishikawa, T., Brazard, J., Usman, A., Sanchez, F., Plaza, P., Martin, M., Falciatore, A., Todo, T., Bouget, F., and C, C. B. (2009). Characterization of two members of the cryptochrome/photolyase family from ostreococcus tauri provides insights into the origin and evolution of cryptochromes. *Plant, Cell and Environment*, 33:1624–1626.

Hiyama, H., Iavarone, A., and Reeves, S. (1998). Regulation of the cdk inhibitor p21 gene during cell cycle progression is under the control of the transcription factor E2F. *Oncogene*, 16(12):1513–1523.

Honkela, A., Girardot, C., Gustafson, E., Liu, Y., Furlong, E., Lawrence, N., and Rattray, M. (2010). Model-based method for transcription factor target identification with limited data. *Proceedings of the National Academy of Sciences*, 107(17):7793–7798.

Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L., Mendes, P., and Kummer, U. (2006). Copasi — a complex pathway simulator. *Bioinformatics*, 22(24):3067–74.

Huang, W., Pérez-García1, P., Pokhilko, A., Millar, A., Antoshechkin, I., Riechmann, J., and Mas, P. (2012). Mapping the core of the arabidopsis circadian clock defines the network structure of the oscillator. *Science*, 336:75–9.

Iacus, S. (2008). *Simulation and Inference for Stochastic Differential Equations*. Springer, Berlin.

Jaakkola, T. (2001). Tutorial on variational approximation methods. In Opper, M. and Saad, D., editors, *Advanced mean field methods: theory and practice*. MIT Press, Cambridge, MA.

Jazwinski, A. (1970). *Stochastic processes and filtering theory*. Academic Press, New York.

Johannes, M. and Polson, N. (2003). Mcmc methods for continuous-time financial econometrics. In Aït-Sahalia, Y. and Hansen, L., editors, *Handbook of Financial Econometrics, Vol 1: Tools and Techniques*. North-Holland, Amsterdam.

Keseler, I., Bonavides-Martinez, C., Collado-Vides, J., Gama-Castro, S., Gunsalus, R., Aaron Johnson, D., Krummenacker, M., Nolan, L., Paley, S., Paulsen, I., et al. (2009). EcoCyc: a comprehensive view of Escherichia coli biology. *Nucleic Acids Research*, 37(1):D464–D470.

Khanin, R., Vinciotti, V., and Wit, E. (2006). Reconstructing repressor protein levels from expression of gene targets in E. coli. *Proceedings of the National Academy of Sciences*, 103(49):18592–18596.

Kitano, H. (2002). Computational systems biology. *Nature*, 420(6912):206–210.

Kloeden, P. and Platen, E. (1992). *Numerical solution of stochastic differential equations*. Springer, Berlin.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models*. MIT Press.

Koller, D., Friedman, N., Getoor, L., and Taskar, B. (2007). Graphical models in a nutshell. In Getoor, L. and Taskar, B., editors, *Introduction to Statistical Relational Learning*. MIT Press, Cambridge, MA.

Komorowski, M., Finkenstädt, B., and Rand, D. (2010). Using a single fluorescent reporter gene to infer half-life of extrinsic noise and other parameters of gene expression. *Biophysical Journal*, 98(12):2759–2769.

Kügler, P. (2012). Moment fitting for parameter inference in repeatedly and partially observed stochastic biological models. *PLoS One*, 7(8):e43001.

Kullback, S. and Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86.

Langevin, P. (1906). Sur la théorie du mouvement brownien. *Comptes Rendues*, 146:530.

Lappalainen, H. and Miskin, J. (2000). Ensemble learning. In Girolami, M., editor, *Advances in Independent Component Analysis*. Springer.

Lawrence, N., Girolami, M., Rattray, M., and Sanguinetti, G., editors (2010). *Learning and inference in computational systems biology*. MIT Press, Cambridge, MA.

Lawrence, N., Sanguinetti, G., and Rattray, M. (2007). Modelling transcriptional regulation using gaussian processes. In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, pages 785–792.

Lèbre, S., Becq, J., Devaux, F., Stumpf, M. P., and Lelandais, G. (2010). Statistical inference of the time-varying structure of gene-regulation networks. *BMC Systems Biology*, 4(1):130.

Lee, C., Ferreon, J., Ferreon, A., Arai, M., and Wright, P. (2010). Graded enhancement of p53 binding to CREB-binding protein (CBP) by multisite phosphorylation. *Proceedings of the National Academy of Sciences*, 107(45):19290–19295.

Lee, C. H., Kim, K. H., and Kim, P. (2009). A moment closure method for stochastic reaction networks. *Journal of Chemical Physics*, 130(13):134107.

Liao, J., Boscolo, R., Yang, Y., Tran, L., Sabatti, C., and Roychowdhury, V. (2003). Network component analysis: reconstruction of regulatory signals in biological systems. *Proceedings of the National Academy of Sciences*, 100(26):15522–15527.

Liepe, J., Barnes, C., Cule, E., Erguler, K., Kirk, P., Toni, T., and Stumpf, M. (2010). Abc-sysbio—approximate bayesian computation in python with gpu support. *Bioinformatics*, 26(14):1797–9.

Liu, M., Durfee, T., Cabrera, J., Zhao, K., Jin, D., and Blattner, F. (2005). Global transcriptional programs reveal a carbon source foraging strategy by Escherichia coli. *Journal of Biological Chemistry*, 280(16):15921–15927.

Lotka, A. J. (1910). Contribution to the theory of periodic reactions. *Journal of Physical Chemistry*, 14(3):271–274.

MacKay, D. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge.

Macneil, L. T. and Walhout, A. J. M. (2011). Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome Research*, 21(5):645–57.

Mangan, S. and Alon, U. (2003). Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences*, 100(21):11980–11985.

Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., DREAM5 Consortium, Kellis, M., Collins, J. J., and Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. *Nature Methods*, 9(8):796–804.

McLean, S., Bowman, L. A. H., Sanguinetti, G., Read, R. C., and Poole, R. K. (2010). Peroxynitrite toxicity in *Escherichia coli* K-12 elicits expression of oxidative stress responses, and protein nitration and nitrosylation. *Journal of Biological Chemistry*, 285:20724–20731.

Miller, C., Schwalb, B., Maier, K., Schulz, D., Dümcke, S., Zacher, B., Mayer, A., Sydow, J., Marcinowski, L., Dölken, L., Martin, D. E., Tresch, A., and Cramer, P. (2011). Dynamic transcriptome analysis measures rates of mrna synthesis and decay in yeast. *Molecular Systems Biology*, 7(458).

Milner, P., Gillespie, C. S., and Wilkinson, D. J. (2013). Moment closure based parameter inference of stochastic kinetic models. *Statistics and Computing*, 23(2):287–295.

Morant, P., Thommen, Q., Pfeuty, B., Vandermoere, C., Corellou, F., Bouget, F., and Lefranc, M. (2010). A robust two-gene oscillator at the core of ostreococcus tauri circadian clock. *Chaos*, 20(4):045108.

Mullis, K. and Faloona, F. (1987). Specific synthesis of dna in vitro via a polymerase-catalyzed chain reaction. *Methods in Enzymology*, 155:335–350.

Murugan, R. and Kreiman, G. (2012). Theory on the coupled stochastic dynamics of transcription and splice-site recognition. *PLoS Computational Biology*, 8(11):e1002747.

Neal, R. (1993). Probabilistic inference using markov chain monte carlo methods. *Technical Report*, pages CRG–TR–93–1.

Oates, C. J., Hennessy, B. T., Lu, Y., Mills, G. B., and Mukherjee, S. (2012). Network inference using steady-state data and goldbeter-koshland kinetics. *Bioinformatics*, 28(18):2342–2348.

Ocone, A., Millar, A. J., and Sanguinetti, G. (2013). Hybrid regulatory models: a statistically tractable approach to model regulatory network dynamics. *Bioinformatics*, 7(29):910–916.

Ocone, A. and Sanguinetti, G. (2011). Reconstructing transcription factor activities in hierarchical transcription network motifs. *Bioinformatics*, 27(20):2873–2879.

Ocone, A. and Sanguinetti, G. (2013). A stochastic hybrid model of a biological filter. In Bortolussi, L., Bujorianu, M., and Pola, G., editors, *Hybrid Autonomous Systems, EPTCS 124*, pages 100–108.

Ogasawara, H., Ishida, Y., Yamada, K., Yamamoto, K., and Ishihama, A. (2007). Pdhr controls the respiratory electron transport system in escherichia coli. *Journal of Bacteriology*, 189(15):5534–5541.

O'Hagan, A. (2003). Hsss model criticism. In Green, P. J., Hjort, N. L., and Richardson, S., editors, *Highly Structured Stochastic Systems*. Oxford University Press, Oxford.

O'Neill, J., van Ooijen, G., Dixon, L., Troein, C., Corellou, F., Bouget, F., Reddy, A., and Millar, A. (2011). Circadian rhythms persist without transcription in a eukaryote. *Nature*, 469:554–8.

Opper, M., Ruttor, A., and Sanguinetti, G. (2010). Approximate inference in continuous time gaussian-jump processes. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel,

R. S., and Culotta, A., editors, *Advances in Neural Information Processing Systems 23*, pages 1831–1839.

Opper, M. and Sanguinetti, G. (2008). Variational inference for markov jump processes. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *Advances in Neural Information Processing Systems 20*.

Opper, M. and Sanguinetti, G. (2010). Learning combinatorial transcriptional dynamics from gene expression data. *Bioinformatics*, 26(13):1623–1629.

Opper, M. and Winther, O. (2001). From naive mean field theory to the tap equations. In Opper, M. and Saad, D., editors, *Advanced mean field methods: theory and practice*. MIT Press, Cambridge, MA.

Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D., and van Oudenaarden, A. (2002). Regulation of noise in the expression of a single gene. *Nature genetics*, 31(1):69–73.

Parisi, G. (1988). *Statistical Field Theory*. Addison Wesley, New York.

Partridge, J., Sanguinetti, G., Dibden, D., Roberts, R., Poole, R., and Green, J. (2007). Transition of Escherichia coli from aerobic to micro-aerobic conditions involves fast and slow reacting regulatory components. *Journal of Biological Chemistry*, 282(15):11230–11237.

Pedraza, J. M. and van Oudenaarden, A. (2005). Noise propagation in gene networks. *Science*, 307(5717):1965–9.

Pokhilko, A., Fernández, A., Edwards, K. D., Southern, M. M., Halliday, K. J., and Millar, A. J. (2012). The clock gene circuit in arabidopsis includes a repressilator with additional feedback loops. *Molecular Systems Biology*, 8(196).

Polager, S. and Ginsberg, D. (2009). p53 and E2f: partners in life and death. *Nature Reviews Cancer*, 9(10):738–748.

Preis, T., Schneider, J., and Stanley, H. E. (2011). Switching processes in financial markets. *Proceedings of the National Academy of Sciences*, 108(19):7674–8.

Prost, S., Lu, P., Caldwell, H., and Harrison, D. (2006). E2F regulates DDB2: consequences for DNA repair in Rb-deficient cells. *Oncogene*, 26(24):3572–3581.

Ptashne, M. and Gann, A. (2002). *Genes and signals*. Cold Harbor Spring Laboratory Press, New York.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Rasmussen, C. and Williams, C. (2006). *Gaussian processes for machine learning*. MIT Press, Cambridge, MA.

Reinker, S., Altman, R. M., and Timmer, J. (2006). Parameter estimation in stochastic biochemical reactions. *Systems Biology (Stevenage)*, 153(4):168–78.

Renshaw, E. (1991). *Modelling biological populations in space and time*. Cambridge University Press, Cambridge.

Risken, H. (1984). *The Fokker-Planck Equation*. Springer, Berlin.

Robert, C. (2001). *The Bayesian Choice*. Springer, Paris.

Rogers, S., Khanin, R., and Girolami, M. (2007). Bayesian model-based inference of transcription factor activity. *BMC Bioinformatics*, 8(Suppl 2):S2.

Rolfe, M. D., Ocone, A., Stapleton, M. R., Hall, S., Trotter, E. W., Poole, R. K., Sanguinetti, G., and Green, J. (2012). Systems analysis of transcription factor activities in environments with stable and dynamic oxygen concentrations. *Open Biology*, 2(7):120091.

Ruttor, A. and Opper, M. (2009). Efficient statistical inference for stochastic reaction processes. *Physical Review Letters*, 103(23):230601.

Ruttor, A., Sanguinetti, G., and Opper, M. (2010). Approximate inference for stochastic reaction processes. In Lawrence, N., Girolami, M., Rattray, M., and Sanguinetti, G., editors, *Learning and inference in computational systems biology*. MIT Press, Cambridge, MA.

Sabatti, C. and James, G. (2006). Bayesian sparse hidden components analysis for transcription regulation networks. *Bioinformatics*, 22(6):739–746.

Sanguinetti, G., Lawrence, N., and Rattray, M. (2006). Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities. *Bioinformatics*, 22(22):2775–81.

Sanguinetti, G., Ruttor, A., Opper, M., and Archambeau, C. (2009). Switching regulatory models of cellular stress response. *Bioinformatics*, 25(10):1280–1286.

Schultz, D., Jacob, E. B., Onuchic, J., and Wolynes, P. (2007). Molecular level stochastic model for competence cycles in bacillus subtilis. *Proceedings of the National Academy of Sciences*, 104(45):17582–7.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–42.

Shahrezaei, V., Ollivier, J., and Swain, P. (2008). Colored extrinsic fluctuations and stochastic gene expression. *Molecular Systems Biology*, 4(196).

Stathopoulos, V. and Girolami, M. (2012). Markov chain monte carlo inference for markov jump processes via the linear noise approximation. *Philosophical Transactions of the Royal Society A: Physical, Mathematical and Engineering Sciences*, 371(61984):20110541.

Stimberg, F., Ruttor, A., and Opper, M. (2012). Bayesian inference for change points in dynamical systems with reusable states - a chinese restaurant process approach. *Journal of Machine Learning Research - Proceedings Track 22*, pages 1117–1124.

Stimberg, F., Ruttor, A., Opper, M., and Sanguinetti, G. (2011). Inference in continuous-time change-point models. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F. C. N., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 24*, pages 2717–2725.

Stumpf, M., Balding, D., and Girolami, M., editors (2011). *Handbook of Statistical Systems Biology*. Wiley, Chichester, UK.

Süel, G. M., Garcia-Ojalvo, J., Liberman, L. M., and Elowitz, M. B. (2006). An excitable gene regulatory circuit induces transient cellular differentiation. *Nature*, 440(7083):545–50.

Swain, P., Elowitz, M., and Siggia, E. (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences*, 99(20):12795–800.

Tebaldi, T., Re, A., Viero, G., Pegoretti, I., Passerini, A., Blanzieri, E., and Quattrone, A. (2012). Widespread uncoupling between transcriptome and translatome variations after a stimulus in mammalian cells. *BMC Genomics*, 13(220).

Thommen, Q., Pfeuty, B., Morant, P.-E., Corellou, F., Bouget, F., and Lefranc, M. (2010). Robustness of circadian clocks to daylight fluctuations: hints from the picoeucaryote ostreococcus tauri. *PLoS Computational Biology*, 6(11):e1000990.

Tian, T., Xu, S., Gao, J., and Burrage, K. (2007). Simulated maximum likelihood method for estimating kinetic rates in gene expression. *Bioinformatics*, 23(1):84–91.

Titsias, M. K., Lawrence, N. D., and Rattray, M. (2009). Efficient sampling for gaussian process inference using control variables. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 21*, pages 1681–1688.

Toner, D. L. and Grima, R. (2013). Molecular noise induces concentration oscillations in chemical systems with stable node steady states. *Journal of Chemical Physics*, 5(138):055101.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. H. (2009). Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202.

Troein, C., Corellou, F., Dixon, L., van Ooijen, G., O'Neill, J., Bouget, F., and Millar, A. (2011). Multiple light inputs to a simple clock circuit allow complex biological rhythms. *The Plant Journal*, 66(2):375–85.

Ukai-Tadenuma, M., Yamada, R. G., Xu, H., Ripperger, J., Liu, A. C., and Ueda, H. R. (2011). Delay in feedback repression by cryptochrome 1 is required for circadian clock function. *Cell*, 144(2):268–81.

Van Kampen, N. (1981). *Stochastic processes in physics and chemistry*. North-Holland, Amsterdam.

Vesely, F. (2001). *Computational physics: an introduction*. Springer, Berlin.

Vogel, C., Abreu Rde, S., Ko, D., Le, S. Y., Shapiro, B. A., Burns, S. C., Sandhu, D., Boutz, D. R., Marcotte, E. M., and Penalva, L. O. (2010). Sequence signatures and mrna concentration can explain two-thirds of protein abundance variation in a human cell line. *Molecular Systems Biology*, 6(400).

Volterra, V. (1926). Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. *Mem. R. Accad. Naz. dei Lincei*, 2:31–113.

Vyshemirsky, V. and Girolami, M. (2008a). Bayesian ranking of biochemical systems models. *Bioinformatics*, 24(6):833–839.

Vyshemirsky, V. and Girolami, M. (2008b). Biobayes: a software package for bayesian inference in systems biology. *Bioinformatics*, 24(17):1933–4.

Wallace, E., Benayoun, M., van Drongelen, W., and Cowan, J. D. (2011). Emergent oscillations in networks of stochastic spiking neurons. *PLoS One*, 6(5):e14804.

Wang, J. and Tian, T. (2010). Quantitative model for inferring dynamic regulation of the tumour suppressor gene p53. *BMC Bioinformatics*, 11(36).

Wilkinson, D. (2011). *Stochastic modelling for systems biology*. Chapman and Hall/CRC Press, London.

Yuen, K. (2010). Relationship between the hessian and covariance matrix for gaussian random variables. pages 257–262. Wiley Online Library, Singapore.

Zechner, C., Ruess, J., Krenn, P., Pelet, S., Peter, M., Lygeros, J., and Koeppl, H. (2012). Moment-based inference predicts bimodality in transient gene expression. *Proceedings of the National Academy of Sciences*, 109(21):8340–5.